

# A Multiple Regression Project

---

## KEYWORDS:

Teaching;  
Diet.

*Roger Johnson*

Carleton College, Northfield, Minnesota, U.S.A.

## Summary

The number of calories in a serving of a food item may be determined from the amount of fat, protein, and carbohydrates it contains. Students can uncover this relationship by collecting food data and then performing a multiple regression.

---

## ◆INTRODUCTION◆

---

TO MINIMIZE the risk of disease due to diet, a number of dietary recommendations have been made. The U.S. National Cancer Institute (1987) and Research Council (1989), for instance, suggest that no more than 30% of the calories we consume consist of fat, and that no more than 10% of the calories we consume consist of simple carbohydrates (sugars). What follows is a project which allows students to determine the above percentages for a given food item from a regression model which they build using food data that they collect. In the process of computing these percentages students are enlightened about misleading advertising claims on the labels of food products.

---

## ◆MODEL BUILDING◆

---

To start the project, have students collect the following (rounded) data from several food items of interest to them: the number of calories, the number of grams of fat, the number of grams of protein, and the number of grams of carbohydrates (each per serving). Students might work in groups of two or three to collect this information from grocery stores, (fast-food) restaurants, and/or their homes. They should disregard food items that do not list contents in grams or do not round to the nearest gram (e.g. a food item that contains "less than 2 gms" of fat). It is useful to have students include at least one food item which advertises how little fat it contains .an example being "2% Milk", and at least one food item which indicates what portion of its carbohydrates are simple carbohydrates (sugars) such as a breakfast cereal. A sample data set is given in Table 1.

Most of my students are initially unaware of the relation

$$\text{Calories} = 9(\text{Fat}) + 4(\text{Protein}) + 4(\text{Carbohydrates})$$

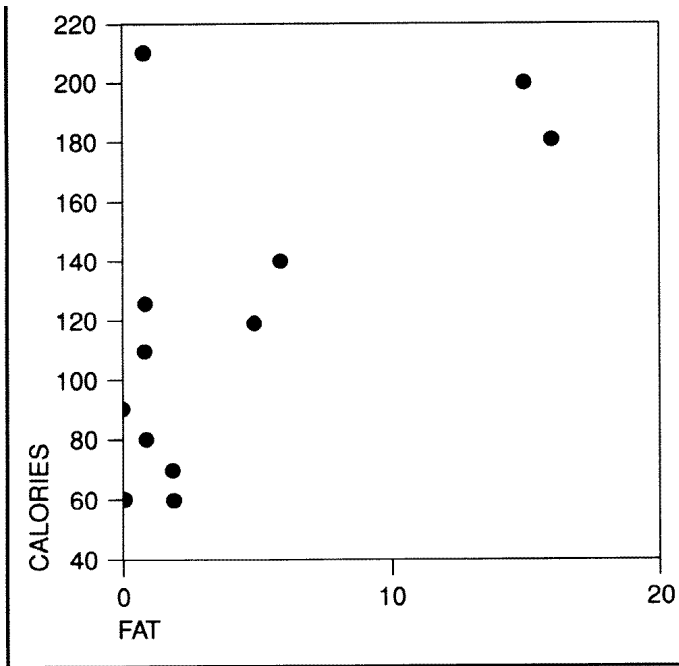
(see Brody (1987), for example) where Fat, Protein, and Carbohydrates (whether simple or complex) are given in grams and the coefficients have units of calories per gram. Students may uncover the above relation by performing a multiple regression. First, to see how the variables Fat, Protein, and Carbohydrates enter the model they can plot Calories against each of one of these individual variables. The pattern in each of these three plots is roughly linear . see Figures 1, 2, 3. Consequently, it makes sense to suppose that

$$\text{Calories} = a + b(\text{Fat}) + c(\text{Protein}) + d(\text{Carbohydrates})$$

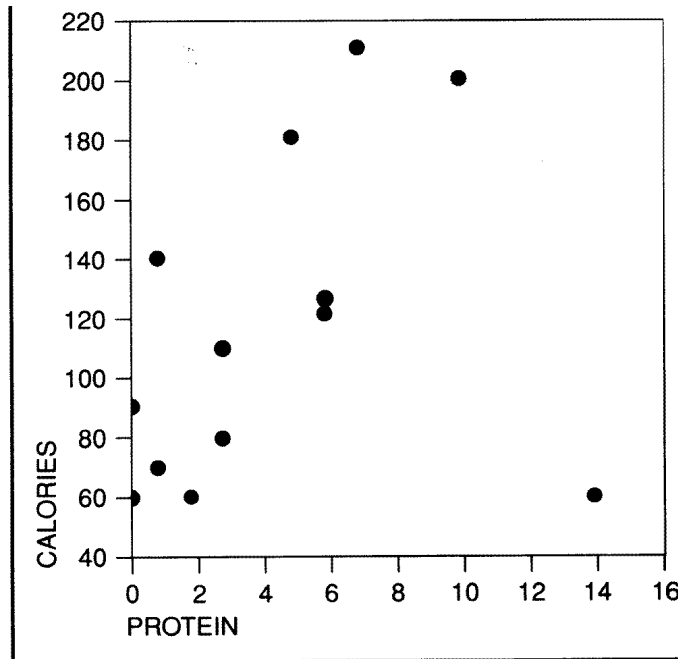
for some a, b, c, d. Next, to show why the model should have a zero constant term the instructor could, for instance, ask what the above model predicts for water, which has no fat, protein, or carbohydrates. Finally, when the zero constant is agreed upon, students can use statistical software to regress Calories on Fat, Protein, and Carbohydrates. (With Minitab, for instance, one can use the regress command with the noconstant subcommand.) Regression through the origin with the data in Table I yields the fitted model:

$$\text{Calories} = 8.89 (\text{Fat}) + 4.27 (\text{Protein}) + 3.98 (\text{Carbohydrates})$$

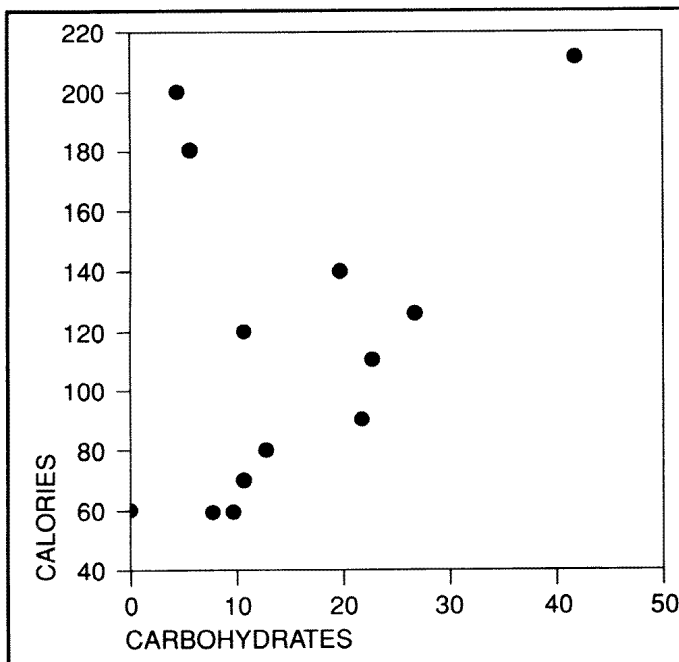
The residual plot in Figure 4 shows that the fit is quite good. Here, the typical size of a residual (the s.e.) is 6.9 calories. It should be noted that in using computer output for this problem that any probabilities (e.g.  $p$  values) are not necessarily accurate. Such probability calculations are done



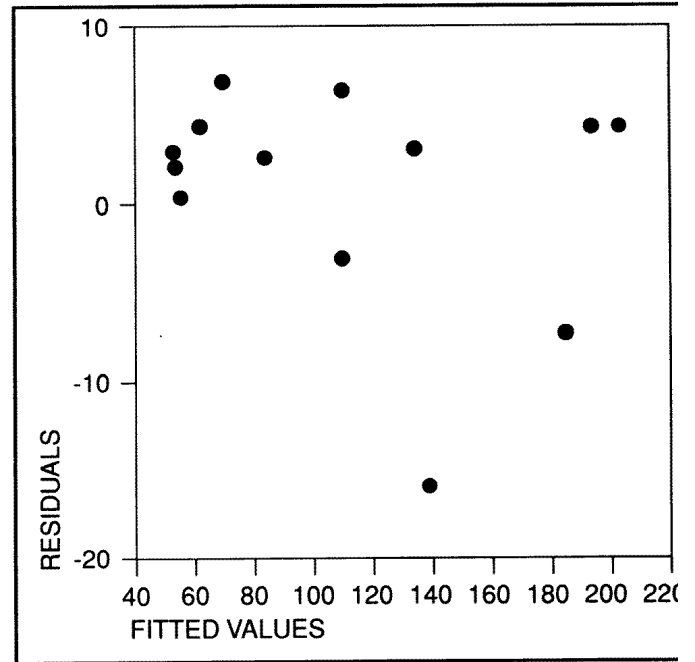
**Fig. 1.** Calories against fat.



**Fig. 2.** Calories against protein.



**Fig. 3.** Calories against carbohydrates.



**Fig. 4.** Residuals against fitted values.

assuming that the calorie errors (reported calories - actual calories) are independent and normally distributed with a common variance. Although independent the calorie errors in our problem are presumably uniformly distributed on either  $[-.5, .5]$  (the baked beans) or  $[-5, 5]$  (the other foods). Regardless of the distributional properties of the errors, however, we are guaranteed that the above multiple regression model minimizes the sum of squared residuals.

Those students that have only been exposed to simple linear regression might initially consider fitting calories to just fat, just protein, or just carbohydrates. The fitted models in these cases are given in Table 2. The typical size of a residual indicates the inadequacy of simple linear regression for our problem of estimating calories and can be used to motivate the need for using several independent variables in the regression model.

**Table 1.** Food data

| Food Item          | Calories | Fat (g) | Protein (g) | Carbo hydrates (g) |
|--------------------|----------|---------|-------------|--------------------|
| 2% Milk            | 120      | 5       | 6           | 11                 |
| Honey Nut Cheerios | 110      | 1       | 3           | 23 (10 sugars)     |
| Plain M & M's      | 140      | 6       | 1           | 20                 |
| Mixed Nuts         | 180      | 16      | 5           | 6                  |
| Tuna               | 60       | 0       | 14          | 0                  |
| Baked Beans        | 126      | 1       | 6           | 27                 |
| Animal Crackers    | 70       | 2       | 1           | 11                 |
| Peanut Butter      | 200      | 15      | 10          | 5                  |
| Chicken Soup       | 60       | 2       | 2           | 8                  |
| Apple Juice        | 90       | 0       | 0           | 22                 |
| Macaroni           | 210      | 1       | 7           | 42                 |
| Instant Coffee     | 60       | 2       | 0           | 10                 |
| Bread              | 80       | 1       | 3           | 13                 |

**Table 2.** Simple linear regressions through the origin.

| Model                           | Typical residual size (s.e.) |
|---------------------------------|------------------------------|
| Calories = 14.74 (fat)          | 85.3                         |
| Calories = 16.28 (protein)      | 84.3                         |
| Calories = 5.69 (carbohydrates) | 71.8                         |

### ◆ ANALYSIS ◆

Noting that each of the three summands in the multiple regression model fitted above have calories for units, the percentage of calories due to fat is approximately

$$100\% (8.89)(\text{grams of fat})/(\text{calories}),$$

the percentage of calories due to protein is approximately

$$100\% (4.27)(\text{grams of protein})/(\text{calories}),$$

and the percentage of calories due to carbohydrates is approximately

$$100\% (3.98)(\text{grams of carbohydrates})/(\text{calories}).$$

Table 3 lists these percentages for the food items in Table 1 (because of rounding errors the row percentages may not add to exactly 100%).

**Table 3.** Percentages of calories due to fat, protein and carbohydrates.

| Food item          | Fat | Protein | Carbo hydrates |
|--------------------|-----|---------|----------------|
| 2% Milk            | 39% | 22%     | 38%            |
| Honey Nut Cheerios | 8%  | 11%     | 81%            |
| Plain M&M's        | 39% | 3%      | 58%            |
| Mixed Nuts         | 76% | 11%     | 13%            |
| Tuna               | 0%  | 100%    | 0%             |
| Baked Beans        | 6%  | 18%     | 76%            |

|                 |     |     |      |
|-----------------|-----|-----|------|
| Animal Crackers | 27% | 6%  | 67%  |
| Peanut Butter   | 68% | 22% | 10%  |
| Chicken Soup    | 31% | 15% | 55%  |
| Apple Juice     | 0%  | 0%  | 100% |
| Macaroni        | 4%  | 15% | 81%  |
| Instant Coffee  | 30% | 0%  | 69%  |
| Bread           | 12% | 17% | 70%  |

Table 3 and extensions of it to include a more complete food listing may be used, for example, by individuals wishing to limit the percentage of calories due to fat in their diet. One cannot have foods that have over 30% of their calories due to fat in "too great" a quantity if they desire this percentage bound to hold for their entire diet. The column for carbohydrates in Table 3 could be further divided into percentage of calories due to simple carbohydrates and the percentage of calories due to complex carbohydrates. In connection with the breakfast cereal Honey Nut Cheerios in our data set, for instance, note

100% (3.98) (10 grams sugars)/(110 calories) = 36% of the calories come from simple carbohydrates. Consequently, as 81% of the calories are due to (total) carbohydrates (see Table 3) 45% of the calories are due to complex carbohydrates.

Students should be asked to reconcile any apparent differences between the advertised food claims they saw with their computations. Note, for example, that "2% Milk" derives 39% of its calories from fat. The advertised percentages in cases such as these are not false, they are simply computed with respect to weight or volume - not with respect to calories.

Our multiple regression model, of course, can now be used to estimate the percentage of calories due to the three food components for foods not used in the model-building process. McDonald's (1991) "McLean Deluxe Sandwich", for example, advertises a "lean beef patty [that] contains 9% fat before cooking." This claim, made in terms of weight, sounds quite healthy. But according to a McDonald's spokesperson the patty has 130 calories and 7 grams of fat so that 48%. or nearly half of the calories in the patty are due to fat. Additional fat, of course, may be added during the cooking process!

### References

- Brody, Jane (1987), *Jane Brady's Nutrition Book*, Bantam Books, New York, N.Y., p. 102.
- McDonald's Corporation (1991), Nutritional Information Center, Oak Brook, IL 60521, (708) 575-3663.
- National Cancer Institute (1987), *Diet, Nutrition, and Cancer Prevention: A Guide to Food Choices*, National Institutes of Health Publication Number 87-2878, U.S. Government Printing Office, Washington, D.C.
- National Research Council (1989), *Diet and Health: Implications for Reducing Chronic Disease Risk*, National Academy Press, Washington, D.C.