Project Title: **Where do all those Calories come from? (Part Two)**
Due Date: **Wednesday, May 10**

Purpose: Explore numerical issues that arise while solving linear systems, especially systems that are composed of nearly equivalent equations.

Teams: Your teams should be the same as for the first project since you will use the data you collected for that project. **Each student must submit their own write-up the project**. Please indicate the name(s) of your partner(s) in your write-up. Teams are limited to a maximum of three people.

Overdetermined systems

Several questions could be asked about our approach and results, such as "what was the purpose of choosing different food groups to work with?", "why were only three foods from each group chosen?", and "why were the results conflicting or inconsistent?" Another question that we can ask after-the-fact is "how could we get more accurate answers given that the data we have access to seems to be inaccurate?" We now turn our attention to trying to answer this last question. Our approach will be to pool data so that we can use many equations to find the correct values for three variables.

A system is *overdetermined* if there are more equations than variables. Overdetermined systems are often (but not necessarily) inconsistent. Any two nonparallel lines in a plane will intersect at a unique point. If there are three or more lines, however, chances are that there will not be a single intersection point for all three lines in which case the system of equations for these lines will have no solution.

Sometimes we expect that an inconsistent linear system *should* have a solution in the sense that if all of our data is accurate we would expect to be able to find a solution. In solving our nutrition problem it would seem reasonable to expect that as the amount of available data we use to find a solution increases we would be able to more accurately compute the Calories due to fat, carbohydrates, and protein. Instead, we obtain an overdetermined system of many equations with three unknowns which will almost certainly be inconsistent. The issue, of course, is that the data we have is not perfect, and so we need a way to find the "best possible" answer given the data we have.

Least Squares

Section 6.5 of our text presents *Least Squares* as an approach to "solve" an inconsistent linear system. The goal is to find the vector $\hat{\mathbf{x}}$ that makes the difference $\mathbf{b} - A\mathbf{x}$ (called the *residual*) as small as possible. The *norm* of a vector $\mathbf{v} \in \mathbb{R}^n$ is defined as $||\mathbf{v}|| = \sqrt{\mathbf{v}^T\mathbf{v}}$ and computes the magnitude (i.e. length) of the vector. (See Section 6.1 for more information.) While we may not be able to find a vector $\mathbf{x}$ to solve $A\mathbf{x} = \mathbf{b}$, we can find the vector $\hat{\mathbf{x}}$ that makes $||\mathbf{b} - A\mathbf{x}||$ as small as possible.

The method of Least Squares leads to a system of equations called the *normal equations*. Assume that A is an $m \times n$ matrix with $m > n$ and $\mathbf{b}$ is an $m \times 1$ vector. Then the normal equations are

$$A^T A\hat{\mathbf{x}} = A^T \mathbf{b}. \tag{1}$$

The solution $\hat{\mathbf{x}}$ of this system is exactly the vector that minimizes $||\mathbf{b} - A\mathbf{x}||$ because it is the orthogonal projection of $\mathbf{b}$ onto Col A.

Assignment

Please provide clear answers and responses by following the directions below.

1. For the first project you collected data from nine different foods and used this to solve three distinct systems. Now combine the data to generate a system $A\mathbf{x} = \mathbf{b}$ with a 9×3 matrix A and 9×1 vector $\mathbf{b}$, then create and solve the normal equations $A^T A\hat{\mathbf{x}} = A^T \mathbf{b}$ to find $\hat{\mathbf{x}}$. Compare this solution to the individual solutions you found in the first project.

2. Compute the residual vector $\mathbf{r} = \mathbf{b} - A\hat{\mathbf{x}}$ and its norm $||\mathbf{r}||$.

3. Compute the norm of the error $||\hat{\mathbf{x}} - \mathbf{x}^*||$, where $\mathbf{x}^* = [9, 4, 4]^T$.

4. Solve the normal equations using the matrix $A^T A$ and vector $A^T \mathbf{b}$ which were constructed from the entire class data set for all food types:

$$A^T A = \begin{bmatrix} 39918 & 67273 & 12352 \\ 67273 & 193151 & 29186 \\ 12352 & 29186 & 6755 \end{bmatrix}, \qquad A^T \mathbf{b} = \begin{bmatrix} 639529 \\ 1469088 \\ 249857 \end{bmatrix}$$

5. Compute the residual vector $\mathbf{r} = \mathbf{b} - A\hat{\mathbf{x}}$ and its norm $||\mathbf{r}||$ using $\hat{\mathbf{x}}$ found in the previous step. How does this residual compare with the one you found in Step 2?

6. Compute the norm of the error $||\hat{\mathbf{x}} - \mathbf{x}^*||$, where $\mathbf{x}^* = [9, 4, 4]^T$. How does does this error compare with the one you found in the Step 3?