Note: this is a two page problem

One can derive the Linear Least Squares Formula using a slightly more sophisticated version of finding the minima of a function which you have learned in past calculus classes. Here, we will walk you through the derivation and introduce some important techniques for doing calculus with vectors and matrices.

(a) For a matrix A and vectors x, b, the expression $Ax - b$ yields a vector. Let's call it y. Let $A_{i,j}$ be the element in the ith row and jth column of the matrix A.

(i) Derive an expression for the ith element of y. Hint: Your solution will involve a summation over one of the two indices. Which index are you summing over?

(ii) The l_2 norm can be represented as a function of the dot product of the vector y with itself. To get ready to write this in the form of a summation, expand the following expression: $(\sum_j A_{i,j} x_j)^2$. Hint: you will need to introduce a third index k in your solution.

(iii) Using your solutions to the prior two problems, write the full expression for $y^T y$. Remember that your answer should be a scalar, so you will need to sum over all indices

(b) Let's call the solution above $f(x)$. Now derive an expression for $\frac{\partial f}{\partial x_\mu}$ where μ is a dummy index. Hint: $\frac{\partial x_i}{\partial x_j} = 0$ if $i \neq j$. We can represent this fact using the Kronecker delta function: $\frac{\partial x_i}{\partial x_j} = \delta_{i,j}$. If you have never seen the Kronecker delta before / don't remember what it is, take a moment to look it up.

(c) Your expression is probably a mess. Let's clean it up

(i) The Kronecker delta makes all the elements of a sum zero except for the one where the two indices match. Use this fact to eliminate extra sums in your expression

(ii) The extra index you added in part (a.ii) is a "dummy index." Notice how index k does not appear in the same expression as index j (if it does, you've made a mistake). Because index j and k don't appear together, we can change index k to index j and combine the terms (why can we do this?). Do so and rewrite your simplified expression.

(d) We are going to turn the your expression from part (c.ii) back into matrix form. Recall that $\sum_j A_{i,j}x_j = (Ax)_i$ and that $A_{i,j} = A^T_{j,i}$. Use these two facts to turn your expression for $\frac{\partial f}{\partial x_\mu}$ into matrix form.

(e) Set this expression equal to zero and solve for x

(f) You did it! Your solution to the last part (the part before b) is the Moore-Penrose pseudo-inverse of A. This is an important concept for when we don't have square matrices in our learning problem (which is most of the time). Take a moment to compare your expression for the Moore-Penrose pseudo-inverse with that listed on the Wikipedia page (you're looking for the case wehre A has linearly independent columns), and summarize what you learned from this problem. **If you have already spent an hour or more on this problem, do not redo earlier parts.**