Vector calculus review (2 + 2 + 4 + 4 + 1 + 2 = 15 pts).

1. Let $g(t) = [g_1(t), g_2(t), \ldots, g_p(t)]^T$ and $h(t) = [h_1(t), h_2(t), \ldots, h_p(t)]^T$. I.e., g and h are column vectors of length p such that each component is a function of (a scalar argument) t. Here, the superscript T stands for the "transpose".

   Show that
   $$\frac{\partial g(t)^T h(t)}{\partial t} = g'(t)^T h(t) + g(t)^T h'(t),$$
   where $g'(t) = [g_1'(t), g_2'(t), \ldots, g_p'(t)]^T$ is the column vector of (component-wise) derivatives with respect to t; similarly, for $h'(t)$. *(Hint: use the definition $g(t)^T h(t) = \sum_{i=1}^p g_i(t)h_i(t)$ and the rule for the derivative of the product.)*

   $$\frac{\partial g(t)^T h(t)}{\partial t} = \frac{\partial \sum_{i=1}^p g_i(t)h_i(t)}{\partial t} = \sum_{i=1}^p g_i'(t)h_i(t) + \sum_{i=1}^p g_i(t)h_i'(t) = g'(t)^T h(t) + g(t)^T h'(t)$$

2. Recall that the gradient of a function f of a vector argument $x = (x_1, \ldots, x_p)$ is a column vector of the partial derivatives of f with respect to the components of x, i.e., $\nabla_x f(x) = [\frac{\partial f(x_1,\ldots,x_p)}{\partial x_1}, \ldots, \frac{\partial f(x_1,\ldots,x_p)}{\partial x_p}]^T$.

   Show that if $f(x) = b^T x$, then $\nabla_x f(x) = b$, where b is a constant column vector of length p (does not depend on x). *(Hint: find the jth partial derivative and then "stack" the partials into a column vector.)*
   Write $b^T x = \sum_{i=1}^p b_i x_i$. Since $\frac{\partial \sum_{i=1}^p b_i x_i}{\partial x_j} = b_j$, $\nabla_x b^T x = b$ (definition).

3. Let $f(x) = x^T A x$, where A is a constant $p \times p$ matrix (that does not depend on x). Show that $\nabla_x f(x) = Ax + A^T x$. *(Hint: show first that $\frac{\partial x^T A x}{\partial x_j} = e_j^T A x + e_j^T A^T x$, where e_j is the column vector with all components equal to zero except the jth, which is 1 (this is the jth standard basis vector). Make use the result in 6.1. Notice that $e_j^T A$ is the jth row of A, while Ae_j is the jth column of A. After finding the expression for the jth partial derivative, "stack" them into a column vector.)*
   Recall that in computing a partial derivative with respect to x_j, all variables that do not depend on x_j are treated as constants. Hence $x^T A x$ can be treated as a function of x_j only when the partial derivative with respect to x_j is concerned. We know how to do this from 6.1. Identify $g(t) = x, h(t) = Ax, t = x_j$. Then $g'(t) = e_j$. Since $Ax = \sum_{i=1}^p C_i x_i$, where C_i is the ith column of A, it is seen that $h'(t) = C_j = Ae_j$. Hence, by 6.1,
   $$\frac{\partial x^T A x}{\partial x_j} = e_j^T A x + x^T A e_j = e_j^T A x + e_j^T A^T x,$$
   since $a^T b = b^T a$ for every column vectors a and b of the same length. Now examine the jth element of the vector $Ax + A^T x$, and notice that it is equal to $\frac{\partial x^T A x}{\partial x_j}$.

4. Let Y be a column vector of length n, β be a column vector of length p and X be a matrix of size $n \times p$. Define $S(\beta) = (Y - X\beta)^T(Y - X\beta)$. Use the results in 6.1-6.3 to find the gradient of $S(\beta)$.

   $$(Y-X\beta)^T(Y-X\beta) = Y^TY - Y^TX\beta - \beta^TX^TY + \beta^TX^TX\beta = Y^TY - 2Y^TX\beta + \beta^TX^TX\beta,$$

   where $Y^TX\beta = \beta^TX^TY$ because $a^Tb = b^Ta$ for every column vectors a and b of the same length. Identifying $A = X^TX$ and $b = X^TY$ and applying the results from 6.2-6.3, one obtains that $\nabla_\beta S(\beta) = -2X^TY + 2X^TX\beta$.

5. Suppose the goal is to find a minimizer of S with respect to β. What are the first-order conditions (on the gradient of S with respect to β) in order for $\widehat{\beta}$ to be a minimizer of S? *If you cannot answer for a general p, take $p = 2$. If still in trouble, examine $p = 1$. (Cannot go any lower, sorry.)*
   If $\widehat{\beta}$ is a minimizer of $S(\beta)$, then it is necessary (but not sufficient) that $\nabla_\beta S(\beta) = -2X^TY + 2X^TX\widehat{\beta} = 0$.

6. What are the second-order conditions? If X is of full rank, does the function S have a unique minimizer with respect to β? *If you cannot answer for a general p, take $p = 2$. If still in trouble, examine $p = 1$. (Cannot go any lower, sorry.)*
   If $\widehat{\beta}$ is a minimizer of $S(\beta)$, it is necessary that $\beta^TX^TX\beta \geq 0$ for every $\beta \neq 0$. If X is of full rank, then $\beta^TX^TX\beta = c^Tc = \sum_{i=1}^p c_i^2 = 0$ only if $\beta = 0$, where $c = X\beta$.

   If X is of full rank, the matrix X^TX is invertible. Solving $\nabla_\beta S(\beta) = 0$ for β yields $\widehat{\beta} = (X^TX)^{-1}X^TY$ which is the unique minimizer of $S(\beta)$.