Note on Matrix Derivatives

STA 2101F: Methods of Applied Statistics I 2022

The matrix version of the linear model is

$$y = X\beta + \epsilon,$$

where y and $X\beta$, and ϵ are $n \times 1$ vectors; X is an $n \times p$ matrix and β is a $p \times 1$ vector. To find the least squares estimator we minimize

$$SS(\beta) = (y - X\beta)^T(y - X\beta) = \sum_{i=1}^{n}(y_i - x_i^T\beta)^2,$$

where x_i^T is the ith row of the matrix X. The text says "differentiating with respect to β and setting to zero we find that $\hat{\beta}$ satisfies

$$X^T X\hat{\beta} = X^T y."$$

Since $SS(\beta)$ has a square, we expect that it's derivative has a 2, so mindlessly

$$SS'(\beta) = 2(y - X\beta)^T(-X),$$

but this has the dimension $1 \times p$, and the derivative with respect to a $p \times 1$ vector needs to be $p \times 1$, so something is not quite right. In class I wrote

$$SS(\beta) = y^T Y - \beta^T X^T y - y^T X\beta + \beta^T X^T X\beta, SS'(\beta) = -X^T y - y^T X + 2X^T X\beta,$$

but these dimensions aren't right either, because $y^T X$ is $1 \times p$ and the other terms are $p \times 1$.

So, let's do it really carefully, starting from

$$SS(\beta) = \sum_{i=1}^{n}(y_i - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2.$$

$$\begin{aligned}
\frac{\partial SS(\beta)}{\partial \beta_1} &= -2\sum_{i=1}^{n} x_{i1}(y_i - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}) = 0 \\
&\vdots \\
\frac{\partial SS(\beta)}{\partial \beta_p} &= -2\sum_{i=1}^{n} x_{ip}(y_i - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}) = 0
\end{aligned}$$

leading to

$$\begin{aligned}
\sum_{i=1}^{n} x_{i1} y_i &= \beta_1 \sum_{i=1}^{n} x_{i1}^2 + \ldots \beta_p \sum_{i=1}^{n} x_{i1} x_{ip} \\
\sum_{i=1}^{n} x_{i2} y_i &= \beta_1 \sum_{i=1}^{n} x_{i1} x_{i2} + \ldots \beta_p \sum_{i=1}^{n} x_{ip} x_{i2} \\
&\vdots \\
\sum_{i=1}^{n} x_{ip} y_i &= \beta_1 \sum_{i=1}^{n} x_{i1} x_{ip} + \ldots \beta_p \sum_{i=1}^{n} x_{ip}^2,
\end{aligned}$$

which we can (hopefully) recognize as the set of equations

$$X^T y = X^T X \hat{\beta}, \qquad \text{leading to } \hat{\beta} = (X^T X)^{-1} X^T y.$$

This will be tedious to carry out each time, so it's nice to have some handy rules. These are widely available on the internet and in algebra textbooks, but some books consider vectors to be row vectors, rather than column vectors, and some solutions I found on the internet were just wrong.

The rule we have 'proved' here is, for an $n \times 1$ vector z and an $n \times p$ vector α,

$$\frac{\partial z^T z}{\partial \alpha} = \left(\frac{\partial z}{\partial \alpha} \right)^T z.$$

This assumes that z is a function of α. This can be (slightly) generalized to

$$\frac{\partial z^T W z}{\partial \alpha} = \left(\frac{\partial z}{\partial \alpha} \right)^T W z,$$

assuming that the $n \times n$ matrix W does not depend on α.

When a matrix depends on a single parameter, e.g. $\Sigma = \Sigma(\theta)$, then $\dfrac{\partial \Sigma}{\partial \theta}$ is interpreted as the matrix with entries $\dfrac{\partial \Sigma_{jk}}{\partial \theta}$. If θ is a $p \times 1$ vector then each of these entries is itself a vector, and we have a tensor! (A 3-dimensional array).