Background summary on functions of vectors

A parametric model for a random variable y is expressed as a density for y that depends on one or more unknown parameters.[1] In regression models, the density is for the conditional distribution of y, given some covariates X.

In linear regression, the simplest model for an independent sample of size n is

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

where Y and ϵ are $n \times 1$ vectors, X is an $n \times p$ matrix and β is a $p \times 1$ vector. If we assume ϵ follows as Normal distribution with expected value 0 and variance-covariance matrix $\sigma^2 I$, then the density is

$$f(y \mid X; \beta, \sigma^2) = \left(\frac{1}{\sqrt{(2\pi)}\sigma}\right)^n \exp\{-\frac{1}{2\sigma^2}(y - X\beta)^{\mathrm{T}}(y - X\beta)\},$$

and the log-likelihood function is

$$\ell(\beta, \sigma^2; y, X) = -n\log(\sigma) - \frac{1}{2\sigma^2}(y - X\beta)^{\mathrm{T}}(y - X\beta).$$

Differentiating $\ell(\beta, \sigma^2; y, X)$ with respect to β gives:

$$\frac{\partial}{\partial\beta}\ell(\beta, \sigma^2; y, X) = -\frac{1}{2\sigma^2}\frac{\partial}{\partial\beta}\left\{(y - X\beta)^{\mathrm{T}}(y - X\beta)\right\},$$

and when we set this to zero we have

$$\frac{\partial}{\partial\beta}(y - X\beta)^{\mathrm{T}}(y - X\beta) = 0.$$

The function being differentiated is a scalar, and β is a column vector of length p, so the result is a column vector of length p. The resulting equation is

$$X^{\mathrm{T}}(y - X\beta) = 0, \tag{1}$$

[1]The density might be a probability mass function, if the variable is discrete.

which you can get by looking up formulas for matrix derivatives[2], or by working it out, component by component:

$$\begin{aligned}(y - X\beta)^{\mathrm{T}}(y - X\beta) &= \sum_{i=1}^{n}(y_i - x_i^{\mathrm{T}}\beta)^{\mathrm{T}}(y_i - x_i^{\mathrm{T}}\beta) \\ &= \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - x_{i2}\beta_2 - \ldots - x_{ip}\beta_p)^2.\end{aligned}$$

Then

$$\frac{\partial}{\partial\beta_j}(y - X\beta)^{\mathrm{T}}(y - X\beta) = -2\sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - x_{i2}\beta_2 - \ldots - x_{ip}\beta_p)x_{ij}$$

and we do this calculation for $j = 1, \ldots, p$, so we have p equations in p unknowns:

$$\begin{aligned}\sum_{i=1}^{n}(y_i - x_i^{\mathrm{T}}\beta)x_{i1} &= 0, \\ \sum_{i=1}^{n}(y_i - x_i^{\mathrm{T}}\beta)x_{i2} &= 0, \\ &\vdots \\ \sum_{i=1}^{n}(y_i - x_i^{\mathrm{T}}\beta)x_{ip} &= 0\end{aligned}$$

and I'll leave it to you to check that in matrix notation this is (1).

See p.364 of SM for the calculation of the matrix of second derivatives

$$\frac{\partial^2\ell(\beta, \sigma^2)}{\partial\beta\partial\beta^{\mathrm{T}}},$$

which has (j, k)th element

$$\frac{\partial^2\ell(\beta, \sigma^2)}{\partial\beta_j\partial\beta_k}.$$

In this course there will be a bit more of this type of calculation for other statistical models, but it won't be a main feature of the course, and if we were having in-person tests, I would not ask you to do that calculation on the test.

[2]in which there is no shame, I look them up all the time