

# Data Science Toolbox Question Sheet

## 02.1 Regression

Daniel Lawson

### Block 2

#### Short questions

1. In the univariate case with  $K = 1$ , the correlation between two samples from the same distribution is written

$$\text{Corr}(X_1, X_2) = \frac{\mathbb{E}[(X_1 - \mu)(X_2 - \mu)]}{\sigma_X^2}.$$

Define multivariate correlation in matrix notation.

2. X and Y are found to be correlated, conditional on some additional variables Z. Under what circumstances does this imply that X causes Y?
3. Consider the linear model  $\mathbf{y} = \mathbf{x}\beta + \epsilon$  where  $\mathbf{x}$  is of dimension  $S$ . Using  $(a + bx)^T = (a^T + x^T b^T)$  and  $a = a^T$  when  $a$  is a 1 by 1 matrix (or otherwise), demonstrate that:

$$\text{MSE}(\beta) = \frac{1}{n}(\mathbf{y}^T \mathbf{y} - 2\beta^T \mathbf{x}^T \mathbf{y} + \beta^T \mathbf{x}^T \mathbf{x} \beta).$$

4. Recall that  $\hat{H} = \mathbf{X}((\mathbf{X}^T \mathbf{X})^{-1})\mathbf{X}^T$ . Show that  $\mathbb{I} - \hat{H}$  is symmetric, by computing  $\hat{H}^T$  or otherwise.
5. Compute  $\hat{H}\hat{H}$  to show that it is Idempotent.
6. Discuss the value of unbiasedness of the estimator for  $\hat{\beta}$ , i.e. that  $\mathbb{E}(\hat{\beta}) = \beta$ .
7. Prove that  $\mathbb{E}(\mathbf{y}) = \mathbf{x}\beta$ .
8. Compute  $\text{Var}(\mathbf{y})$  and discuss the circumstances in which it simplifies.