

UNIVERSITY OF TECHNOLOGY, SYDNEY

School of Mathematical Sciences

**Advanced Data Analysis**

**ASSIGNMENT 1**

**Due time and date:** 5pm Monday, 14th March, 2011.

**Submission location:** Drop-box Number 4, near the lifts on Level 15 of Building 1.

1. Random variables  $X$  and  $Y$  have joint density function

$$f_{X,Y}(x, y) = 8xy, \quad 0 < x < y < 1.$$

- (a) Find the conditional density function of  $Y$  given  $X = x$ .  
(b) Determine  $E(Y|X = x)$ .

2. Let  $X_1, \dots, X_n$  be a random sample from the Gamma( $\alpha, \beta$ ) distribution:

$$f_X(x; \alpha, \beta) = \frac{e^{-x/\beta} x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha}, \quad x > 0.$$

Let  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ . Show that the maximum likelihood estimators  $\hat{\alpha}$  and  $\hat{\beta}$  satisfy

$$\hat{\beta} = \bar{X}/\hat{\alpha}.$$

3. Let  $\mathbf{X}$  be a  $p \times 1$  random vector with mean  $E(\mathbf{X}) = \boldsymbol{\mu}$ . Also, let  $\boldsymbol{\Sigma}$  be the covariance matrix of  $\mathbf{X}$ . Show that, for any symmetric  $p \times p$  matrix  $\mathbf{A}$ ,

$$E(\mathbf{X}^T \mathbf{A} \mathbf{X}) = \boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu} + \text{tr}(\mathbf{A} \boldsymbol{\Sigma}).$$

4. Consider least squares fitting of the general linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where  $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ . The vector of fitted values is given by

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

where  $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ .

- (a) Show that  $\mathbf{H}$  is idempotent; i.e.  $\mathbf{H}^2 = \mathbf{H}$ .  
(b) Show that  $\mathbf{H}$  is unaffected by linear transformations of the columns of  $\mathbf{X}$ . That is, show that  $\mathbf{X}_L = \mathbf{X}\mathbf{L}$ , for any invertible matrix  $\mathbf{L}$ , leads to the same expression for  $\mathbf{H}$ .
5. The accompanying graphic shows data on calories versus sodium content for a random sample of sausages. The sausages are made of meat of 3 different type: beef, pork and poultry. A nutritionist is interested in comparing mean calory content of beef sausages with poultry sausages, and pork sausages with poultry sausages — controlling for sodium content. Write down an appropriate regression model for achieving this goal. You may assume that there are no significant interactions.

Hint: Indicator variables (also known as 'dummy' variables) play a role here.

