Exercises 2

1. Read Rice section 14.4.1 on vector valued random variables, mean vector and covariance matrix and answer the following.

   (a) For a linear transformation $\mathbf{Z} = \mathbf{AY} + \mathbf{c}$ where $\mathbf{Y}$ is a random vector, $\mathbf{A}$ is a fixed matrix and $\mathbf{c}$ is a fixed vector (all of matching dimensions for matrix algebra), state how the mean vector $\mathrm{E}(\mathbf{Z})$ and covariance matrix $\mathbf{V}(\mathbf{Z})$ are related to $\mathrm{E}(\mathbf{Y})$ and $\mathbf{V}(\mathbf{Y})$.

   (b) Now refer to equation (2.10) in Section 2.1.3 of the lecture notes. Assuming that the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}$ is non-singular, use the results in part (a) to obtain $\mathrm{E}(\hat{\boldsymbol{\beta}})$ and $\mathbf{V}(\hat{\boldsymbol{\beta}})$. (Note that the first two results in Section 2.1.1(iv) of the notes are special cases.)

   [Also read the remainder of Rice section 14.4, in particular for the proof of the result for E(RSS) in Section 2.1.2(ii) of the notes.]

2. Suppose the model for simple linear regression is written as

$$Y_i = \alpha_0 + \alpha_1(x_i - \overline{x}) + e_i \qquad (i = 1, \ldots, N)$$

   where $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

   (a) Obtain the least squares estimators of α_0 and α_1. (You may obtain these from the matrix form of the normal equations.)

   (b) Obtain the covariance matrix of these least squares estimators under the usual assumptions about the errors.

   (c) Show that

$$\mathrm{RSS} = C_{YY} - \frac{C_{xY}^2}{C_{xx}}$$

   where $C_{xx} = \sum_{i=1}^{N}(x_i - \overline{x})^2, C_{xY} = \sum_{i=1}^{N}(x_i - \overline{x})(Y_i - \overline{Y})$ and $C_{YY} = \sum_{i=1}^{N}(Y_i - \overline{Y})^2$.

   (d) Refer to R output 1 (flow data). If x denotes log depth and y denotes log flow, you are given that $\overline{x} = -0.88686$, $\overline{y} = 0.21475$, $C_{xx} = 1.1482, C_{xy} = 3.1738, C_{yy} = 9.3516$ using the notation above.

   State how the model defined above is related to the model fitted in the R output. Use the algebraic results obtained above and a hand calculator to verify the numerical results for the following in the R output:

   - the least squares estimates of β_0 and β_1,
   - the residual standard error,
   - estimated covariance matrix of $\hat{\beta}_0$ and $\hat{\beta}_1$.

3. Assume a model

$$Y_i = \mathbf{x}_i^T\boldsymbol{\beta} + e_i,\ e_i \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}, i = 1, \ldots, N + 1.$$

   Given that you know $Y_i, \mathbf{x}_i,\ i = 1, \ldots, N,\ \mathbf{x}_{N+1}$ for observation no. $N + 1$, but not Y_{N+1}. Use the theory in 2.1.4 to

   (a) write down a $100(1 - \alpha)\%$ confidence interval for $\mathbf{x}_{N+1}^T\boldsymbol{\beta}$,

   (b) show that $\frac{\mathbf{x}_{N+1}^T\hat{\boldsymbol{\beta}} - \mathbf{x}_{N+1}^T\boldsymbol{\beta} - e_{N+1}}{\sqrt{\hat{\sigma}^2(v+1)}} \sim t_{N-p}$, where $v = \mathbf{x}_{N+1}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{N+1}$,

   (c) using (b), write down a $100(1 - \alpha)\%$ prediction interval for Y_{N+1}.

4. The data of example A in the notes and handout 3 (tree volume data) can be obtained as a text file `treevol.dat` from the course webpage.

   Since the volume of a cylinder is a product of the height, the diameter to the square and a constant, it could actually make sense to assume a multiplicative model, and fitting the log VOL from log HT and log D16 could make sense.

Use R or any other statistics software to fit such a model. Discuss the results and compare them to the models in handout 3.

Compute a predicted value and 95% prediction and confidence intervals for VOL for a tree with D16=10 and HT=100 from the log-transformed model fitted here (you need to take into account the transformation for this). Compare the prediction interval to the one that you get from a model that fits VOL as a linear function of D16 and HT.

Here are some useful R-commands:

```
treevol <- read.table("treevol.dat",header=TRUE)
logHT <- log(treevol$HT)
logD16 <- log(treevol$D16)
logVOL <- log(treevol$VOL)
pairs(cbind(logVOL,logHT,logD16))
lmtreelog <- lm(logVOL ~ logHT + logD16)

newtree <- data.frame(HT=100,D16=10)
lognewtree <- data.frame(logHT=log(newtree$HT),logD16=log(newtree$D16))
predict(lmtreelog,lognewtree,interval=c("confidence"),level=0.95)
predict(lmtreelog,lognewtree,interval=c("prediction"),level=0.95)
```