**STA 410S/2102S**: Practice questions for 1st test

**These do not cover all the topics that may be on the test**

The test will have 4 questions of equal value; you are permitted one aid sheet (8.5 x 11).

1. The following R program, due to Radford Neal, computes the Cholesky decomposition of a symmetric, positive definite matrix.

```
cholesky <- function (A)
{
  if (!is.matrix(A) || nrow(A)!=ncol(A))
  { stop("The argument for cholesky must be a square matrix")
  }

  p <- nrow(A)
  U <- matrix(0,p,p)

  for (i in 1:p)
  {
    if (i==1)
    { U[i,i] <- sqrt (A[i,i])
    }
    else
    { U[i,i] <- sqrt (A[i,i] - sum(U[1:(i-1),i]^2))
    }

    if (i<p)
    { for (j in (i+1):p)
      { if (i==1)
        { U[i,j] <- A[i,j] / U[i,i]
        }
        else
        { U[i,j] <- (A[i,j] - sum(U[1:(i-1),i]*U[1:(i-1),j])) / U[i,i]
        }
      }
    }
  }

  U

}
```

The program is uncommented; please add comments to explain how the program works. Illustrate it on the matrix

$$\begin{pmatrix} 9 & -3 & 6 \\ -3 & 2 & -3 \\ 6 & -3 & 6 \end{pmatrix}$$

2. Consider the linear model

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

where $y$ is $n \times 1$, $X$ is $n \times p$ and has column rank $p$, $\beta$ is a $p \times 1$ vector of unknown coefficients, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ where we assume $\epsilon_i \sim (0, \sigma^2)$ and the $\epsilon_i$ are independent.

   (a) Recall that the least squares estimator $\hat{\beta} = (X^T X)^{-1} X^T y$; from which

   $$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy$$

   where $H$ is called the hat matrix. Show that $H$ is idempotent.

   (b) Show that $\hat{\epsilon}^T \hat{\epsilon} = y^T (I - H) y$, where $\hat{\epsilon} = y - \hat{y}$.

   (c) Show that $\hat{\epsilon}^T \hat{\epsilon}/(n - p)$ is an unbiased estimate of $\sigma^2$.

3. Assume our (very old) computer does floating point arithmetic in which the mantissa has just three decimal digits. Construct an example in which round-off error causes

$$\Sigma(x_i - \bar{x})^2 \neq \Sigma(x_i^2) - n\bar{x}^2.$$

4. The `abbey` dataset contains 31 determinations of nickel content in a rock sample. The values are:

```
> abbey
 [1]   5.2   6.5   6.9   7.0   7.0   7.0   7.4   8.0   8.0   8.0   8.0
[12]   8.5   9.0   9.0  10.0  11.0  11.0  12.0  12.0  13.7  14.0  14.0
[23]  14.0  16.0  17.0  17.0  18.0  24.0  28.0  34.0 125.0
```

Following the book I computed several summary statistics in R, as follows:

```
> mean(abbey)
[1] 16.00645
> median(abbey)
[1] 11
> unlist(hubers(abbey))
       mu         s
11.731514  5.258487
> unlist(hubers(abbey,k=2))
       mu         s
12.351117  6.105222
> unlist(hubers(abbey,k=1))
       mu         s
11.365392  5.567345
> unlist(huber(abbey))
      mu        s
11.55136  4.44780
> mad(abbey)
[1] 4.4478
> IQR(abbey)
[1] 7
```

Explain to a non-statistician why all these estimates of 'mu' are different. Which one would you recommend?

5. A random variable $U$ has the *uniform* distribution on $(0,1)$ ($U \sim U(0,1)$) if its distribution function is

$$\Pr(U \leq u) = \begin{cases} 0 & u < 0 \\ u & 0 \leq u \leq 1 \\ 1 & u > 1 \end{cases}$$

Show that if $X$ has a distribution function $F(x)$, then $Y = F(X)$ follows a uniform distribution on $(0,1)$. Show conversely that if $U \sim U(0,1)$ that $Z = F^{-1}(U)$ has distribution function $F$. Use this to write an R program that simulates a sample of size $n$ from

$$F(x) = 1 - \exp(-\lambda x)$$

using the function `runif` which generates samples from a $U(0,1)$ distribution. Your function should take $n$ and $\lambda$ as input and return a random sample of length $n$.

Here are the topics that we have covered so far:

| | |
|---|---|
| floating point arithmetic | class notes |
| **basics of R**: | |
| data frames, vectors and matrices, subsetting | class notes and §2.1,2,3 (to p.32) |
| **linear regression**: | |
| fitting models in R | class notes, p.144 |
| computation of OLS estimates by QR decomposition | class notes |
| Cholesky decomposition and singular value decomposition | class notes, p.62,63 |
| diagnostics and residual plots | §6.3 |
| robust estimation of location and scale | §5.5 |
| robust regression | §6.5 to p.159 |
| density estimation | §5.6 |
| generalized linear models | handout |
| permutation tests | HW 1 |
| randomized block designs | HW 1 |