THE UNIVERSITY OF NEW SOUTH WALES

DEPARTMENT OF STATISTICS

ASSIGNMENT 2, MATH3811/MATH3911- Statistical Inference

A reminder that assignments count as part of your assessment. Please, write a declaration stating that the assignment is your own work and that you have read and understood the University Rules regarding Student Academic Misconduct.

This assignment must be submitted to your tutor at the **beginning** *of the tutorial (2:00 pm) on Friday, May 27, 2011 at the latest.*

Math3811: Attempt the first four questions. Math3911: Attempt all questions.

1. Suppose $X_1, X_2, \ldots, X_{10}$ are independent and identically distributed random variables from $N(\mu, 4)$ (each with a density $f(x;\mu) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{8}}$).

   a) Show that the joint density of $X_1, X_2, \ldots, X_{10}$ has monotone likelihood ratio in $\sum_{i=1}^{10} X_i$.

   b) Derive the UMP unbiased size $\alpha = 0.05$ test of $H_0 : \mu = 1$ versus $H_1 : \mu \neq 1$.

   c) Express the power function of this test using the cdf of the standard normal and evaluate the power numerically for $\mu = 0, 0.5, 1.5, 2$, and 3. Sketch the power function of the above test for all values of μ on the real axis.

   d) Calculate the density $f_{X_{(2)}}(x)$ of the second order statistic $X_{(2)}$.

2. In a Bayesian setting, we sample n i.i.d. observations $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ from a population with conditional distribution of each single observation being the geometric distribution

$$f_{X_1|\Theta}(x|\theta) = \theta^x(1-\theta), x = 0, 1, 2, \ldots; 0 < \theta < 1.$$

   a) If the prior on Θ is the uniform distribution on (0,1) (that is $\tau(\theta) = 1,\ 0 < \theta < 1$), find the posterior $h(\theta|\mathbf{X} = (x_1, x_2, \ldots, x_n))$.

   b) Determine the Bayes estimator of θ with respect to quadratic loss.
   Hint: For $\alpha > 0$ and $\beta > 0$ the beta function $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$ satisfies $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ where $\Gamma(\alpha) = \int_0^\infty exp(-x)x^{\alpha-1}dx$ is the Gamma function.

3. Important measures in exploratory data analysis are the *skewness* $\gamma_1 = \frac{E(X-E(X))^3}{Var(X)^{3/2}}$ and the *kurtosis* $\gamma_2 = \frac{E(X-E(X))^4}{Var(X)^2} - 3$. One way of estimating them is by using their empirical counterparts $\hat{\gamma_1} = \frac{\sqrt{n}\sum_{i=1}^n (X_i-\bar{X})^3}{(\sum_{i=1}^n (X_i-\bar{X})^2)^{3/2}}$ and $\hat{\gamma_2} = \frac{n\sum_{i=1}^n (X_i-\bar{X})^4}{(\sum_{i=1}^n (X_i-\bar{X})^2)^2} - 3$, respectively.

   a) Similarly to the `mycorr` function from the bootstrap tutorial script, write down two S-PLUS (or R) functions *myskewness* and *mykurtosis* to get the estimates $\hat{\gamma_1}$ and $\hat{\gamma_2}$.

   b) Use the above functions to estimate γ_1 and γ_2 using the variable `time` from the `lung` data set in S-PLUS (this variable records survival times of lungcancer patients from the Mayo clinic in USA). Verify your results by comparing them with the ones obtained when using the S-PLUS in-built functions *skewness* and *kurtosis* and explain the differences (if any).

c) Bootstrap the $\hat{\gamma}_1$ and $\hat{\gamma}_2$ estimators by using $B = 3000$ replicates and report the resulting 95% confidence intervals using the BCa method.

d) For a normal population, theoretical skewness and kurtosis are equal to zero. If the 95% confidence intervals for skewness and kurtosis exclude zero, the normality is in doubt. What is your conclusion about the normality of the `time` variable? Double check your claim graphically by executing the `qqplot` command. Comment. Include your coding, and the output containing the BCa confidence intervals, in your assignment.

4. We simulate an example to demonstrate the unparallelled strength of the LTS regression in isolating outliers when there is some idea about the amount of contamination in the data. If your student number contains the seven numbers XXXXXXX, to generate data that is unique to your student number, you should include the number in the starting seed for random number generation. After setting the initial seed, generate pairs of observations $(x_i, y_i, i = 1, 2, \ldots, 100)$ of which 70% are scattered around the line $y = x + 2$ and 30% are clustered around the point (6,2). Plot the data and superimpose the three regression lines: the least squares line, the default M-estimate line and the default least trimmed squares:

```
>set.seed(round(log(XXXXXXX)))
>x70<-runif(70,0.5,4)
>e70<-rnorm(70,mean=0,sd=0.2)
>y70<-2+x70+e70
>x30<-rnorm(30,mean=6,sd=0.5)
>y30<-rnorm(30,mean=3,sd=0.5)
>x<-c(x70,x30)
>y<-c(y70,y30)
>simuldata<-data.frame(x,y)
>plot(x,y)
>abline(lm(y~x,simuldata));text(6.0,3.0,"LS")
>abline(rreg(x,y));text(5.2,3.0,"M")
>abline(ltsreg.formula(y~x,simuldata));text(4.2,3,"LTS")
```

Execute the commands and print out the resulting graph. Now, (compare the handout about robust regression) in `ltsreg.formula` you are allowed to override the default value of `quan` (the number of residuals included in the least trimmed squares calculations). Modify the instruction about the lts regression by instructing that only 70 residuals be included in the calculation. Redraw the graph with the new LTS regression line replacing the old one, obtain and report the estimated slope and intercept of this new line, edit the graph, comment on it and include the new graph to your assignment.

5. To jackknife an estimator T_n, we calculate the n statistics $T_n^{(i)}, i = 1, 2, \ldots, n$ where $T_n^{(i)}$ is calculated like T_n except that X_i is removed from the sample. Then the jackknife estimator is $JK(T_n) = nT_n - \frac{n-1}{n}\sum_{i=1}^{n} T_n^{(i)}$. Let $X_1, X_2, \ldots, X_n$ be iid uniform in $[0, \theta)$. We know that $T_n = X_{(n)}$ is biased for θ with a bias equal to $-\frac{1}{n+1}\theta$.

a) Express $JK(T_n)$ via order statistics.

b) Show that the bias of $JK(T_n)$ is equal to $-\frac{\theta}{n(n+1)}$ hence indeed the order of the bias of $T_n = X_{(n)}$ has been reduced after jackknifing.

c) Show that $Corr(X_{(n-1)}, X_{(n)}) \to 1/\sqrt{2}$ as $n \to \infty$.