STAT 568 – Assignment 1 – due date is on course outline

For each of the questions which are carried out on R – please e-mail me your R programs. These should be in one file, named `yourname_asst1.R`*, with the various questions clearly separated in the file. I should be able to run them exactly as I receive them, and see the output nicely displayed (see the help files for the* `cat` *function). For these questions your written assignment then need contain only the discussions of your solutions. Also: I will be quite unhappy if you merely recycle my own programs, from the course web site, with only the data sets changed.*

1. Here you will investigate hypothesis testing in a 'canonical form' which greatly simplifies the theory. The method yields the same F-tests as Cochran's Theorem, but in a different way. Suppose the problem begins as follows. You are given a linear model of the data, in which observations $Y_1, ..., Y_n$ are independently and normally distributed, with common variance σ^2, and with means depending on certain predictors $\{x_{ij}\}$. In matrix form, with $\boldsymbol{y} = (Y_1, ..., Y_n)'$, the model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{X}$ is an $n \times p$ matrix of rank $r \leq p$, $\boldsymbol{\theta}$ is a vector of unknown parameters ranging over all of $\mathbb{R}^p$, and $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$. You wish to estimate the parameters by maximum likelihood, and to then carry out the likelihood ratio test of

$$H_0 : \boldsymbol{C}\boldsymbol{\theta} = \boldsymbol{0},$$

where $\boldsymbol{C}$ is a $q \times p$ matrix of rank $q < r \leq p$.

   (a) Consider the following situation, and write it in the manner described above. Specify the matrices $\boldsymbol{X}$ and $\boldsymbol{C}$, and the values of p, r and q.
   There are n subjects in an experiment, n_1 of whom receive a certain drug and $n_2 = n - n_1$ of whom are in a control group and receive nothing. A linear regression model is fitted to the responses Y_i, with one independent variable X – the indicator of the event the the subject receives the treatment. We are to test the hypothesis that the treatment has no effect.
   Parts (b) – (e) are to be done <u>in general</u>, not just in the context of the example in (a), to which you will return in part (f).

   (b) Define $\boldsymbol{\xi} = E[\boldsymbol{y}]$. Show that the problem is equivalent to the following. We observe $\boldsymbol{y} \sim N(\boldsymbol{\xi}, \sigma^2 \boldsymbol{I}_n)$, where $\boldsymbol{\xi}$ lies in a given vector space Π of dimension r. The hypothesis specifies that $\boldsymbol{\xi}$ lies in a particular s-dimensional subspace Π_0 of Π, where $s \leq p - q$.

(c) Recall the Gram-Schmidt method (for instance from the Stat 312 notes); by this one can construct an orthogonal basis $\{\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_s\}$ of Π_0, extend it to an orthogonal basis $\{\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_s; \boldsymbol{\pi}_{s+1}, ..., \boldsymbol{\pi}_r\}$ of Π, and then to an orthogonal basis $\{\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_s; \boldsymbol{\pi}_{s+1}, ..., \boldsymbol{\pi}_r; \boldsymbol{\pi}_{r+1}, ..., \boldsymbol{\pi}_n\}$ of $\mathbb{R}^n$. Let $\boldsymbol{Q}_0$ be the $n \times s$ matrix with columns $\{\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_s\}$, $\boldsymbol{Q}_1$ the $n \times r - s$ matrix with columns $\{\boldsymbol{\pi}_{s+1}, ..., \boldsymbol{\pi}_r\}$, and $\boldsymbol{Q}_2$ the $n \times n - r$ matrix with columns $\{\boldsymbol{\pi}_{r+1}, ..., \boldsymbol{\pi}_n\}$. Then $\boldsymbol{Q}_j'\boldsymbol{Q}_j$ is an identity matrix for each j, and $\boldsymbol{Q}_j'\boldsymbol{Q}_k$ is a zero matrix if $j \neq k$. Now the model states that $\boldsymbol{\xi}$ lies in the column space of $\left(\boldsymbol{Q}_0 \vdots \boldsymbol{Q}_1\right)$, and the hypothesis is that it lies in the column space of $\boldsymbol{Q}_0$. Define

$$\boldsymbol{Q} = \left(\boldsymbol{Q}_0 \vdots \boldsymbol{Q}_1 \vdots \boldsymbol{Q}_2\right),$$

an $n \times n$ orthogonal matrix, and $\boldsymbol{z} = \boldsymbol{Q}'\boldsymbol{y}$. Let $\boldsymbol{\eta}$ be the mean vector. Show that $\boldsymbol{z} \sim N(\boldsymbol{\eta}, \sigma^2 \boldsymbol{I}_n)$, that the model specifies that

$$\boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\eta}_0 \\ \boldsymbol{\eta}_1 \\ \boldsymbol{0} \end{pmatrix} \begin{matrix} \leftarrow s \\ \leftarrow r - s \\ \leftarrow n - r \end{matrix},$$

and that the hypothesis specifies that $\boldsymbol{\eta}_1 = \boldsymbol{0}_{r-s \times 1}$.

(d) Partition $\boldsymbol{z}$ as

$$\boldsymbol{z} = \begin{pmatrix} \boldsymbol{z}_0 \\ \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \end{pmatrix} \begin{matrix} \leftarrow s \\ \leftarrow r - s \\ \leftarrow n - r \end{matrix}.$$

Show that the likelihood ratio test of the hypothesis rejects for large values of

$$F = \frac{\|\boldsymbol{z}_1\|^2 / (r - s)}{\|\boldsymbol{z}_2\|^2 / (n - r)},$$

and that

$$F \sim F_{n-r}^{r-s}\left(\lambda^2 = \frac{\|\boldsymbol{\eta}_1\|^2}{\sigma^2}\right).$$

(e) Define $S(\boldsymbol{\theta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2$. Show that, in terms of the original parameterization,

$$F = \frac{\frac{\min_{H_0} S(\boldsymbol{\theta}) - \min S(\boldsymbol{\theta})}{\nabla df}}{\frac{\min S(\boldsymbol{\theta})}{df\,(mse)}},$$

where the minima are taken with and without the restrictions imposed by the hypothesis, $df\,(mse)$ is the df of the MSE in the unrestricted model, and ∇df

is the difference in the degrees of freedom of the MSEs with and without the hypothesis. Show further that

$$\sigma^2\lambda^2 = \min_{\boldsymbol{t}\in\Pi_0} \|\boldsymbol{\xi} - \boldsymbol{t}\|^2,$$

where $\boldsymbol{\xi}$ is the true mean vector and the minimum is evaluated over mean vectors $\boldsymbol{t}$ as specified by the hypothesis; i.e. is the squared distance from $\boldsymbol{\xi}$ to the closest member of Π_0 (and so is 0 under H_0).

(f) Identify Π, Π_0 and the noncentrality parameter, in the example from (a). As a check, you should notice that the ncp coincides with the F test statistic when the parameters are replaced by estimates.

2. Suppose that $X_1, ..., X_m$ are independent, with $X_i \sim N(\nu_i, 1)$, so that $X^2 = \sum_{i=1}^m X_i^2 \sim \chi_n^2\left(\lambda^2\right)$, with $\lambda^2 = \sum_{i=1}^m \nu_i^2$.

(a) Show that we can assume that $X_1 \sim N(\lambda, 1)$ and that $X_2, ..., X_m \sim N(0, 1)$. [Hint: Let $\mathbf{x}_{m\times 1}$ have elements X_i, and write $X^2 = \|\mathbf{x}\|^2 = \|\mathbf{Q}\mathbf{x}\|^2$ for any orthogonal $\mathbf{Q}$. Choose $\mathbf{Q}$ to have first row $\boldsymbol{\nu}'/\|\boldsymbol{\nu}\|$.]

(b) Use (a) to show that $X^2 \sim X_1^2 + X_{m-1}^2$, where $X_1^2 \sim \chi_1^2\left(\lambda^2\right)$ independently of $X_{m-1}^2 \sim \chi_{m-1}^2$ (central). [This is continued, so as to obtain the density of X^2, on the Stat 575 web site.]

(c) Show that X^2 is 'stochastically increasing in λ', in that the function $P\left(X^2 > c\right)$ is an increasing function of λ, for any $c > 0$. [Hint: Show that X_1^2 has this property, and then condition on X_{m-1}^2.]
Note: The same conditioning approach applies to a singly non-central F_n^m r.v. – it too is stochastically increasing in its ncp, implying that the power of the LR test of $H_0 : \lambda^2 = 0$ – the hypothesis from the previous problem – increases as one moves away from the null hypothesis.

3. From the text: ch. 2 #11. (In all testing situations, state the p-value.)

4. From the text: ch. 2 #12.

5. In the pulp/operator data set, represent the model as a regression model with response

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3,$$

using indicators $X_j = I\left(\text{treatment } j+1\right)$ for $j = 1, 2, 3$. Fit this model to the data. Represent and test the hypothesis of no treatment effects, and verify that the F has the same value as in the anova formulation of the model.

6. Obtain the ANOVA Table 3.19 in R.

7. Obtain the ANOVA Table 3.25 in R, and make the multiple comparisons - which compounds are significantly different at the 5% level? Plot the 95% confidence intervals.

8. From the text: ch. 3 #9. Include a discussion of how blocking might be incorporated.

9. From the text: ch. 3 #33. 'Compare the results' by carrying out a suitable test.

10. From the text: ch. 3 #35. Compare the methods at the 95% level.