

# Homework 1

Due February 1, 2011

**Directions** Please turn-in a hard copy of your R code along with a brief write-up of the solutions (do not submit raw output). Also submit via e-mail (njc23@pitt.edu) a copy of your R code.

Some of the questions below can be answered with very little or no programming. However, write code that outputs the final answer and does not require any additional paper calculations. For example, suppose I ask for how many numbers are greater than 5 in the vector,  $\mathbf{x} = c(1, 9, 2, 8, 10, 12)$ . Do not write  $\mathbf{x} > 5$  and then count the number of TRUEs by hand, instead have R count the number of TRUEs by coding `sum(x>5)`.

- Below is a trivariate probability mass function,  $P(X, Y, Z)$ . Enter the density into R as an array and include the dimension names and element labels. Use the `apply()` function and the `[ ]` to find the following probabilities and distributions.

		X		
Z	Y	10	20	30
0	1	.04	.24	.12
	-1	.01	.06	.03
1	1	.04	.24	.12
	-1	.01	.06	.03

- $P(X = 20, Y = -1, Z = 0)$
  - $P(X \geq 20, Y = -1)$
  - $P(X)$
  - $P(X, Z)$
  - $P(X|Y, Z)$
- Below is a one-way ANOVA model for comparing three treatments with  $n_i$  patients receiving treatment  $i, i = 1, 2, 3$ . Let  $Y_{ij}$  denote the response of the  $j$ th patient receiving the  $i$ th treatment,  $\mu$  be the overall mean response,  $\alpha_i$  the treatment effect and  $\epsilon_{ij}$  the error term. Then,

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

where,  $i = 1, 2, 3$  and  $j = 1, \dots, n_i$  with  $n_1 = 25, n_2 = 30$  and  $n_3 = 20$ . The response depends on the subscripts  $i$  and  $j$ ,

$$Y_{ij} = \begin{cases} 5 & \text{if } i \times j \text{ is even} \\ 15 & \text{if } i \times j \text{ is odd} \end{cases}$$

Enter the design matrix  $\mathbf{X}$  and the response  $\mathbf{Y}$  into R and calculate a least squares estimator of  $\boldsymbol{\beta}$  using the generalized inverse,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^g \mathbf{X}^T \mathbf{Y}.$$

(Hint, use `rep()`).

3. R has a built-in character vector of US State names, `state.name`. Use this character vector and R's character functions to answer the following questions.

- (a) List all the US State names that are more than one word. How many are there?
- (b) What is the longest US State name(s) (including spaces) and how long is it?
- (c) What is the longest single word US State name and how long is it?
- (d) List all the US State names, where all of the upper and lower case "a"s are replaced with a capital "Z".

4. Mrs. Smith is participating in a clinical trial. She started treatment on March 3, 1999 and is expected to have a follow-up visit every 6-months for the next three years.

- (a) Create a vector of the dates when Mrs. Smith is expected to show up for her follow-up visit.
- (b) Mrs. Smith showed up on the following dates, enter these dates into R as Date objects.

9/10/1999, 2/10/2000, 9/16/2000, 2/23/2001, 9/9/2001 3/28/2002

- (c) It is unrealistic to expect Mrs. Smith to show up on the dates from part (a). However, did Mrs. Smith have a follow-up appointment within three weeks of each expected follow-up date?

5. Write an R function that accepts an  $n \times n$  matrix  $\mathbf{A}$  as an argument and returns the trace of  $\mathbf{A}$ ,  $\text{trace}(\mathbf{A}) = \sum_{i=1}^n a_{ii}$ . Include code that verifies that the function's argument is indeed a square matrix. Also include comments that describe the function's purpose, argument and output.

The following code generates a list of 25 matrices of different dimensions. Apply your trace function to each matrix in this list.

```
matrix.list <- lapply(1:25, FUN=function(x) {matrix(1:x^2, nrow=x, ncol=x)})
```