This document provides optional exercises to help prepare students for the 3 credit UNC Chapel Hill course *BCB720 Introduction to Statistical Modelling* taught by Dr. Will Valdar, typically every fall. These exercises are given in the format of the actual homeworks set in BCB720, with point values for each question indicating the relative amount of effort required. For those actual homeworks, students are asked to submit their answers electronically as a PDF file created using Latex, Rmarkdown or another R/Latex system (eg, knitr). It is recommended (although not required) that before students take the BCB720 course, they first make sure they can answer the questions below in the format requested. The questions include mathematics used through the course, namely, calculus (about the level of Calculus 1), elementary matrix algebra, and miscellaneous algebra, along with some practice exercises in writing using Latex math.

For many students, answering these questions will require looking up things they may have forgotten about or perhaps never knew; and as a result, completing the exercises may take some time. Indeed, **many students have *started* the course unable to complete all of these exercises and ultimately completed the course successfully**. Nonetheless, any time spent on this will save time later on, since all of the material here is used in the course and learning it in advance will greatly ease the burden of the first few homeworks, as well as make the lectures easier to follow.

For working with R, it is common for students to use the freely available RStudio program. RStudio is also used by some students for writing documents in R/Markdown, a plain text-like markup format that can include Latex and R code and can be used to produce pdf and Word format files. Other students, however, run R and write Latex through alternative programs. Students are encouraged to show initiative and seek out these alternatives themselves. (For interest, the instructor uses R on the Unix command line and writes Latex using the terminal program Emacs, but acknowledges that this solution will not suit all tastes.)

**Question 1: Algebra, calculus, and notation** (10 points)

Solve the following questions without the aid of a computer (eg, do not use Google, R, etc) and provide a (latex-formatted) solution. Throughout, $\log x$ refers to the natural logarithm of $x$.

(a) Let $f(x) = 9x - x^2$.

i. Let $K$ be area under the curve $f(x)$ in the range $0 \leq x \leq 9$. Use integration to show that $K = \frac{3^5}{2}$. (1)

ii. Let $g(x) = K^{-1} f(x)$ such that the area under $g(x)$ integrates to one. Show that (1)

$$\int_0^9 x\, g(x)\, dx \;= 4.5\,.$$

iii. Show using differentiation that $\text{argmax}_x\, g(x)$, the value of $x$ that maximizes $g(x)$, is 4.5. (1)

iv. Show using differentiation that it is also the case that $\text{argmax}_x \log g(x) = 4.5$. (1)

(b) Let $f(x) = e^{-x^2/2}$. If $g(x) = e^{9x} f(x)^2 = e^y$. Show that $y = 9x - x^2$. (1)

(c) If $f(x) = \log x^4$, then what is $f'(x)$, the differential of $f(x)$ with respect to $x$? (1)

(d) If $f(x) = x^4(1-x)^6$, and $g(x) = \log f(x)$, then find $g'(x)$, the differential of $g(x)$ with respect to $x$. [Note: It should not be necessary to multiply out the product.] (2)

(e) If $f(\alpha) = \prod_{i=1}^n e^{x_i}$, then what is the missing term $A$ in $\log f(\alpha) = \sum_{i=1}^n A$ ? (1)

(f) Let $g(\alpha) = \sum_{i=1}^n \theta y_i$. Differentiate $g(\theta)$ with respect to $\theta$. (1)

**Question 2: Vectors and matrices in R** (12 points)

In R, construct the variables `y`, `X` and `theta` as

$$\mathbf{y} = \begin{bmatrix} 6.2 \\ 3.5 \\ 5.2 \\ 7 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 0.9 & 0.3 & 0.5 \\ 1 & 0 & 0.3 & 0.5 \\ 1 & 0.9 & 0 & 0.5 \\ 1 & 0.9 & 0.3 & 0 \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} 2 \\ 4 \\ 3 \\ 80 \end{bmatrix},$$

and then evaluate the expressions (or demonstrate the results) below using R. [Hint: remember to use matrix multiplication (`%*%`)].

(a) Find $\mathbf{X}^{\mathrm{T}}$, the transpose of $\mathbf{X}$. (1)

(b) Show that $\mathbf{y}^{\mathrm{T}}\mathbf{1} = 21.9$, where $\mathbf{1}$ is a conformable vector of ones. (1)

(c) Show that $\mathbf{X}\boldsymbol{\theta} = (46.5, 42.9, 45.6, 6.5)^{\mathrm{T}}$. (1)

(d) Show that $|\mathbf{X}| = -0.135$, (ie, the determinant of $\mathbf{X}$). (1)

(e) Find $\mathbf{X}^{-1}$, the inverse of $\mathbf{X}$. (1)

(f) Find $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\boldsymbol{\theta}$. (1)

(g) Show that $\mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{y} = 271.942$. (1)

(h) $\mathrm{tr}(\mathbf{X}) = 1$, where $\mathrm{tr}(\cdot)$ denotes the trace of its argument. (1)

(i) If $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$, show that $\boldsymbol{\beta} = (3.3, 3, \frac{10}{3}, -1.6)^{\mathrm{T}}$. (2)

(j) If $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, show that $\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon} = (-40.3, -39.4, -40.4, 0.5)^{\mathrm{T}}$. (2)

**Question 3: Latex Equations** (11 points)

Copy out the statements below using Latex for the equations. To do this, you may compose the statement in Latex itself or in something that uses Latex such as R/Markdown. In either case, show both the output (eg, $\alpha$) and the code that generated it (eg, `$\alpha$`); the TAs should be able to paste the equation code into either Latex or conversion tools such as www.latex2png.com and get the same formula. Partially correct answers will receive partial points. Note that some equations are formatted as in-line and some are formatted as display, and that by convention bold lower case letters are vectors, bold upper case letters are matrices, and `\mathbf` and `\boldsymbol` are best used for roman and greek letters respectively. Although this material will be covered in the course, this exercise is purely about practicing writing latex; **it is not necessary to understand the content at this stage**.

(a) A common measure of diversity used in multiple fields, and throughout bioinformatics, is the so-called "information entropy" defined as $S = -\sum_{j=1}^{J} p_j \log_2 p_j$, where $p_j$ is the proportion of items belonging to the $j$th category for $j \in \{1, \ldots, J\}$. (1)

(b) In the linear model setup, a response $y_i$ of individual $i = 1, \ldots, n$ that is predicted by two variables $x_{1,i}$ and $x_{2,i}$ can be modeled as if it were generated as (2)

$$y_i = \mu + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i \,,$$

where $\mu$ is the intercept, $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ is individual-specific noise, and $\beta_1$ and $\beta_2$ are effects.

(c) (1)

$$\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \end{bmatrix} \quad \text{and} \quad \mathbf{x}_i = \begin{bmatrix} 1 \\ x_{1,i} \\ x_{2,i} \end{bmatrix}$$

(d) Using vector notation, $y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i$. (1)

(e) In matrix form, for all observations at once, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. (1)

(f) The likelihood for all observations is (2)

$$f(y_1, \ldots, y_n | \mathbf{x}_1, \ldots, \mathbf{x}_n, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 \right\}$$

(g) In matrix form, (2)

$$f(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$$

(h) The effects vector $\boldsymbol{\beta}$ can be estimated as $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ (1)