Problem Set 4: Managing Data

You may not use any downloaded packages. Solutions that require a discussion or an explanation should be type-written in a 12-point font and submitted in class—do not include any R code that is not referenced in your solution. See the instructions in Problem Set 2 regarding the submission of R code.

The following functions, which have not been discussed in class, may or may not be useful for this problem set:

`all, any, identical, sample, which`

1. **[3 pts]** Suppose the following plain-text file is saved in your working directory.

wonderPets.txt
```
Linny,Tuck,
Ming-Ming,Ollie,
The Visitor
```

Explain why the imported data contain two empty strings:

```
> scan("wonderPets.txt", what = character(), sep = ",")
Read 7 items
[1] "Linny"       "Tuck"        ""            "Ming-Ming"
[5] "Ollie"       ""            "The Visitor"
```

2. **[3 pts]** Recall the data on the effect of caffeine on the performance of a simple task, finger tapping. The first three lines of the file containing the data are shown below.

caffeine.txt
```
  0 100 200
242 248 246
245 246 248
```

```
> caffeine <- read.table("caffeine.txt", header = TRUE)
> head(caffeine, n = 2)
   X0 X100 X200
1 242  248  246
2 245  246  248
```

The column names look puzzling. Explain.

3. Regression analysis is a statistical method for modeling the relationship between a variable of interest, called the response variable, and one or more related variables, called the covariates. Specifically, the variation in the response variable Y is partly explained by the variation in a set of covariates $X_1, \ldots, X_p$:

$$Y = f(X_1, \ldots, X_p) + \text{noise},$$

where the (unknown) function f is called the regression function. In a linear regression through the origin,

$$f(X_1, \ldots, X_p) = \beta_1 X_1 + \cdots + \beta_p X_p,$$

where $\beta_1, \ldots, \beta_p$ are the parameters to be estimated from sample data. Let

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} x_{11} & \ldots & x_{1p} \\ x_{21} & \ldots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{np} \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}.$$

Here $\mathbf{y}$ is a vector of n observations on Y and $\mathbf{X}$ is a design matrix with x_{ij} representing the ith observation on the jth covariate X_j.

(a) **[3 pts]** The file `ps04p3.RData` contains two objects, `y` and `X`, saved in a binary format. Load the objects into R and determine the class and the number of elements in each of `y` and `X`.

(b) **[3 pts]** Use vectorized operations to name the columns of `X` as `x1`, `x2`, and so on.

(c) **[3 pts]** The objects `y` and `X` correspond to $\mathbf{y}$ and $\mathbf{X}$, respectively. An estimator of $\boldsymbol{\beta}$, denoted $\hat{\boldsymbol{\beta}}$, is the solution to the equation

$$(\mathbf{X}^{\mathsf{T}}\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathsf{T}}\mathbf{y}, \tag{1}$$

where the superscript T denotes the transpose of a matrix. Most textbooks express the solution as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\,\mathbf{y} \tag{2}$$

when $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ is invertible. Compute $\hat{\boldsymbol{\beta}}$ using equation (2) and *only* the following functions and operators: `t` for transpose, `solve` for matrix inversion, and `%*%` for matrix multiplication. Save the solution as `beta1` and report the time it takes to compute $\hat{\boldsymbol{\beta}}$.

(d) **[3 pts]** Compute $\hat{\boldsymbol{\beta}}$ using equation (2) and *only* the following functions: `solve` for matrix inversion and `crossprod` for matrix multiplication. Save the solution as `beta2` and report the time it takes to compute $\hat{\boldsymbol{\beta}}$.

(e) **[3 pts]** Compute $\hat{\boldsymbol{\beta}}$ using equation (1) and *only* the following functions: `solve` for solving a system of linear equations (rather than matrix inversion) and `crossprod` for matrix multiplication. Save the solution as `beta3` and report the time it takes to compute $\hat{\boldsymbol{\beta}}$.

(f) **[3 pts]** Use the function `all.equal` with the optional argument `tol = 1e-12` to determine if the solutions `beta1`, `beta2`, and `beta3` are nearly equal.

(g) **[3 pts]** Compare the time it takes to compute $\hat{\boldsymbol{\beta}}$ in (c)–(e) and explain the differences, if any.

(h) **[3 pts]** Convert the object `X` into a data frame and repeat (c). Explain the result.

4. Throughout this problem, specify only the relative pathname and not the absolute pathname. The file `ps04p4.txt` contains the names of the students enrolled in MATH 267P.

   (a) **[3 pts]** Import the data into a vector named `name` and create an email address for each student as follows. The general format of an email address is `username@domain`. For each student, `username` is the name of the student in lowercase, with a period separating the first name and the last name if a last name is provided; and `domain` is `ponyville.edu`. Name the vector that contains the email addresses as `email`.

   (b) **[3 pts]** Randomly assign the students to groups of three for the first problem set, using uppercase letters (A, B, and so on) to label the groups. Create a data frame named `ps1grps` that contains the names of the students, their email addresses, and the groups they are assigned to. Name the columns as `name`, `email`, and `ps1`.

   (c) **[3 pts]** Export the data in `ps1grps` to a plain-text file named `ps1grps.txt` in the following format:

      - Use the column names as the header line.
      - There should be no quotes or row names.
      - Use the comma as a separator.

   (d) **[3 pts]** Import the data in `ps1grps.txt` to a data frame named `ps1grps.in`. Verify that the data frames `ps1grps.in` and `ps1grps` are exactly the same.