B. Kedem, Spring 2007.

STAT 741 TAKEHOME EXAM

Problem 1

Consider the vector of measurements $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})'$ taken on subject i, and assume the following model,

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{1}$$

where $\mathbf{X}_i$ and $\mathbf{Z}_i$ are nonrandom $n_i \times p$ and $n_i \times q$ design or covariate matrices, respectively, $\boldsymbol{\beta}$ is a p–dimensional parameter of *fixed effects* common to all subjects, $\mathbf{b}_i$ is q–dimensional subject–specific (nonrandom) parameter considered nuisance, and $\boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2 \mathbf{I})$ is the n_i–dimensional vector representing noise or measurement error. Notice that the covariance matrix $\sigma^2 \mathbf{I} = \sigma^2 \mathbf{I}_{n_i}$ is $n_i \times n_i$ and its elements do not depend on i. Denote by $f_i(\mathbf{y}_i | \mathbf{b}_i, \boldsymbol{\beta}, \sigma^2)$ the density of $\mathbf{Y}_i$.

1. Show that $\mathbf{Z}_i' \mathbf{Y}_i$ is sufficient for $\mathbf{b}_i$, and obtain the conditional pdf $f_i(\mathbf{y}_i | \mathbf{Z}_i' \mathbf{y}_i, \boldsymbol{\beta}, \sigma^2)$.

2. Argue that one may try to get inference about $\boldsymbol{\beta}, \sigma^2$ is by maximizing the conditional likelihood $\prod_{i=1}^{N} f_i(\mathbf{y}_i | \mathbf{Z}_i' \mathbf{y}_i, \boldsymbol{\beta}, \sigma^2)$ but that this may not be a good idea.

3. Assume that the $\mathbf{b}_i$ are random, each having the same distribution $Q(\cdot)$. Explain how to get inference about $\boldsymbol{\beta}, \sigma^2$ and Q.

4. A simpler alternative assumes that the random effects are normally distributed in which case model (1) is referred to as *linear mixed–effects model*, where $\mathbf{X}_i$, $\mathbf{Z}_i$, and $\boldsymbol{\beta}$ are as above, but $\mathbf{b}_i \sim \mathcal{N}_q(\mathbf{0}, \mathbf{D})$, and $\boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \boldsymbol{\Sigma})$ is the n_i–dimensional error. Here $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_i$ is $n_i \times n_i$ covariance matrix whose components do not depend on i. It is assumed that the $\boldsymbol{\epsilon}_i$ and $\mathbf{b}_i$ are independent. Obtain inference about all the parameters, and show how to test $\mathrm{H}_0 : \mathbf{C}\boldsymbol{\beta} = \boldsymbol{\beta}_0$ versus $\mathrm{H}_0 : \mathbf{C}\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$.

Problem 2

Let $X \sim \mathrm{b}(\mathrm{n}, \pi)$, and consider the distribution of $X|X > 0$. X is said to have a truncated binomial distribution.

1. Argue that the distribution of $X|X > 0$ is a tilted or distorted distribution, and obtain its form.

2. Data were collected on the number of girls and boys in 28 families in the form of pairs (No. girls, No. boys):

   10 10 11 11 11 11 11 11 11 11 11 20 20 20 21 21 21 21 12 12 30 31 31 13 13 40 41 14

   Since there is at least one girl in each family, do the data suggest the number of girls follow a truncated binomial distribution $X|X > 0$, where $X \sim \mathrm{b}(\mathrm{n}, 1/2)$?

Problem 3

The data file "mtRainyDays" gives the number of Rainy days (response), Max Temperature, and Mean heating Degree. Analyze the data using appropriate modeling. Justify your approach and your final model. This should include a careful examination of the residuals, and appropriate testing.