Stat 771, Fall 2011: Homework 2

Due Wednesday, February 23

1. The file `insulin.dat` contains longitudinal data from a study on $m = 36$ rabbits; 12 rabbits were randomly assigned to each of 3 groups: group 1 rabbits received the standard insulin mixture, group 2 rabbits received a mixture containing 1% less protamine than the standard, and group 3 rabbits received a mixture containing 5% less protamine. Rabbits were injected with the assigned mixture at time 0, and blood sugar measurements taken on each rabbit at the time of injection (time 0) and 0.5, 1.0, 1.5, 2.5, and 3.0 hours post-injection. Each data record in the file `insulin.dat` represents a single observation; the columns of the data set are (1) rabbit number, (2) hours (time), (3) response (blood sugar level), and (4) insulin group (1, 2, or 3).

   You may need to transform the data into a form suitable for the various SAS PROCs (or R functions) to accomplish the following.

   (a) Obtain a profile (spaghetti) plot for each of the three insulin groups with a LOESS smooth superimposed on top. Describe what you see in terms of patterns of blood sugar reduction. Does there seem to be a difference among groups? Are the estimated mean functions (LOESS) approximately parallel?

   (b) Obtain the sample covariance and sample correlation matrix for each insulin group, as well as the pooled sample covariance matrix and corresponding estimated correlation matrix under the assumption of a common covariance matrix. Based these matrices, do you think the assumption of a common covariance matrix for each group is reasonable? Why or why not?

   (c) What covariance model(s) do you think are appropriate for these data? Why?

   (d) Fit the univariate repeated measures model discussed in Chapter 5 for these data:

   $$Y_{hlj} = \mu + \tau_l + \gamma_j + (\tau\gamma)_{lj} + b_{hl} + e_{hlj},$$

   where $h = 1, \ldots, 12$ rabbits in each group $l = 1, 2, 3$, observed over time points $j = 1, 2, 3, 4, 5, 6, 7$. As usual, $b_{hl} \overset{iid}{\sim} N(0, \sigma_b^2)$ independent of $e_{hlj} \overset{iid}{\sim} N(0, \sigma_e^2)$. Report the (correct) ANOVA table and estimates of σ_b and $\rho = \sigma_b^2/(\sigma_b^2 + \sigma_e^2)$. Interpret your estimate of ρ. Is there a significant time by treatment interaction here? We will quantify group differences later on when we discuss the general linear model.

2. Elston and Grizzle (1962) present repeated measurements of ramus (jaw) bone height on a cohort of 20 boys over an 18 month period; the data in `ramus.txt`. Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_{20}$ be the 20 4-dimensional vectors containing the ramus height. Obtain the sample mean vector $\bar{\mathbf{Y}}$ and covariance matrix $\widehat{\mathbf{\Sigma}}$ for these data (e.g. SAS PROC CORR). Also obtain a scatterplot matrix of all pairings of observation times as in class (e.g. PROC SGSCATTER).

   (a) What is happening with the sample correlation between pairs of measurements as the time between observations increases?

   (b) Which of the five types of covariance matrices might be appropriate for these data based on $\widehat{\mathbf{\Sigma}}$?

3. *Repeated measures versus cross-sectional data*

Consider testing a new sleep aid: the amount of sleep obtained in hours is recorded for subject i at baseline (time zero), Y_{i1}, and then the next night the amount of sleep is recorded (on the same subject) after taking the drug Y_{i2}.

Say

$$\begin{bmatrix} Y_{i1} \\ Y_{i2} \end{bmatrix} \sim N_2\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right), \quad \text{i.e. } \mathbf{Y}_i \sim N_2\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right).$$

Let

$$\bar{\mathbf{Y}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Y}_i = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} Y_{i1} \\ \frac{1}{n}\sum_{i=1}^{n} Y_{i2} \end{bmatrix} = \begin{bmatrix} \bar{Y}_1 \\ \bar{Y}_2 \end{bmatrix}.$$

Take it as fact that

$$\bar{\mathbf{Y}} \sim N_2\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n\right).$$

Let's say that $\sigma_1 = \sigma_2 = 1$ and $n = 100$.

A natural estimator of $\mu_2 - \mu_1$ is $d = \bar{Y}_2 - \bar{Y}_1$, i.e. $d = \begin{bmatrix} 1 & -1 \end{bmatrix}\bar{\mathbf{Y}}$ (a matrix times a multivariate normal random vector). What is the variance of this estimator when $\rho = 0.9$, i.e. high correlation among the repeated measures? **Hint**: what is the distribution of $d = \bar{Y}_1 - \bar{Y}_2$?

Now $\rho = 0$ corresponds to independent observations, which occurs with *cross-sectional* data, i.e. two totally different groups of 100 individuals, 100 of which are monitored for baseline sleep, and 100 which are monitored for drug-aided sleep. What is the variance of d when $\rho = 0$? What implications does this have for repeated measures versus cross-sectional data?

4. (Extra credit). *Expectation of χ_p^2 random variable*

Recall that

$$\textbf{Fact: } E(\mathbf{Y}'\mathbf{A}\mathbf{Y}) = tr(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}.$$

(a) Let $\mathbf{Z} \sim N_p(\mathbf{0}, \mathbf{I}_p)$, equivalent to $Z_1, \ldots, Z_p \overset{iid}{\sim} N(0,1)$. Show

$$X = \mathbf{Z}'\mathbf{Z} = \sum_{i=1}^{p} Z_i^2.$$

The random variable X has a χ_p^2 distribution. Use the **Fact** above to show $E(X) = p$.

(b) Now let $\mathbf{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let $\boldsymbol{\Sigma}^{1/2}$ be the unique symmetric matrix such that $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}$, and let $\boldsymbol{\Sigma}^{-1/2}$ be the symmetric inverse of $\boldsymbol{\Sigma}^{1/2}$. Show that $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu}) \sim N_p(\mathbf{0}, \mathbf{I}_n)$. **Hint**: $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2}$. Also recall that if $\mathbf{A}$ is symmetric then $\mathbf{A}' = \mathbf{A}$.

(c) Finally, show $(\mathbf{Y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \boldsymbol{\mu}) \sim \chi_p^2$.