

ST 732, HOMEWORK 1, SPRING 2007

The first few exercises are meant to familiarize you with some operations that we will summarize using matrix notation throughout the course. Use of SAS to carry out the analyses we will discuss requires familiarity with these matrix operations. Future homeworks will not involve matrix algebra problems like these except where relevant.

1. Let Y_1 and Y_2 be random variables with means μ_1 and μ_2 , variances σ_1^2 and σ_2^2 , and covariance σ_{12} , respectively. Let c_1 and c_2 be constants.

(a) **Starting with the definition of variance** in equation (3.2) of the notes and results given in Chapter 3, show that

$$\text{var}(c_1Y_1 + c_2Y_2) = c_1^2\sigma_1^2 + c_2^2\sigma_2^2 + 2c_1c_2\sigma_{12}.$$

(b) Let

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}.$$

Write down the covariance matrix $\mathbf{\Sigma}$ of \mathbf{Y} .

(c) Let

$$\mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}.$$

Verify that $\text{var}(\mathbf{c}'\mathbf{Y}) = \mathbf{c}'\mathbf{\Sigma}\mathbf{c}$ (bottom of p. 45 of the notes) by evaluating each side of this equation and showing that they are equivalent.

The covariance matrix of a linear combination of elements of a random vector is of routine interest in longitudinal analysis.

2. Let Y_1 and Y_2 be as in Problem 1. Let

$$\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}.$$

(a) Find $\mathbf{C}\mathbf{Y}$ and $\text{var}(\mathbf{C}\mathbf{Y})$ using results from Problem 1 and the **definition of covariance**.

(b) By doing the matrix multiplication $\mathbf{C}\mathbf{\Sigma}\mathbf{C}'$, show that $\text{var}(\mathbf{C}\mathbf{Y}) = \mathbf{C}\mathbf{\Sigma}\mathbf{C}'$ as on p. 46 of the notes, where $\mathbf{\Sigma}$ is the matrix you found in Problem 1(b).

The covariance matrix of a general linear function of elements of a random vector is of routine interest in longitudinal analysis.

3. (a) Let b and e be **statistically independent** random variables. Show that $E(be) = E(b)E(e)$.

(b) Suppose we have the following statistical model:

$$Y_{ij} = \mu + b_i + e_{ij},$$

where $i = 1, \dots, m$ and $j = 1, \dots, n$, $b_i \sim \mathcal{N}(0, \sigma_b^2)$, $e_{ij} \sim \mathcal{N}(0, \sigma_e^2)$, b_i and e_{ij} are **statistically independent** for each i and j , and e_{ij} and e_{ik} are statistically independent for any two values

$j, k = 1, \dots, n$. Find the variance of Y_{ij} and the covariance and correlation between any two values Y_{ij} and Y_{ik} , $j \neq k$.

(c) Let $n = 3$, and define

$$\mathbf{Y}_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \end{pmatrix}.$$

Find the covariance matrix of \mathbf{Y}_i , and express it in terms of a (3×3) identity matrix and the (3×3) matrix \mathbf{J} whose elements are all equal to 1.

(d) Show that the covariance matrix in (c) is of the form of one of the popular models in Section 4.4 of the notes by finding the associated correlation matrix and identifying to which popular model it corresponds.

We will see that this covariance matrix plays a special role in a few weeks.

4. Suppose that \mathbf{Y} is a random vector with mean vector $\boldsymbol{\mu}$, where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \end{pmatrix}.$$

(a) Suppose that we are interested in the null hypothesis $H_0 : \mu_1 - \mu_2 = 0, \mu_1 - \mu_3 = 0, \mu_1 - \mu_4 = 0, \mu_1 - \mu_5 = 0$, which may be written alternatively as

$$H_0 : \begin{pmatrix} \mu_1 - \mu_2 \\ \mu_1 - \mu_3 \\ \mu_1 - \mu_4 \\ \mu_1 - \mu_5 \end{pmatrix} = \mathbf{0},$$

where $\mathbf{0}$ is a (4×1) vector of zeros.

Give an appropriate matrix \mathbf{L} so that H_0 may be written in the form $H_0 : \mathbf{L}\boldsymbol{\mu} = \mathbf{0}$.

(b) Now find an appropriate matrix \mathbf{L} corresponding to the null hypothesis $H_0 : \mu_1 - \mu_2 = 0, \mu_2 - \mu_3 = 0, \mu_3 - \mu_4 = 0, \mu_4 - \mu_5 = 0$. Do the hypotheses in (a) and (b) address the same issue or different issues? Explain.

(c) Find the matrix \mathbf{U} such that you can express the hypothesis in (b) in the form $H_0 : \boldsymbol{\mu}'\mathbf{U} = \mathbf{0}'$; note that now $\mathbf{0}'$ is a (1×4) vector, so H_0 is being expressed as a **row** vector instead of a column vector.

(d) Now suppose we are interested in the null hypothesis $H_0 : \mu_1 - (\mu_2 + \mu_3 + \mu_4 + \mu_5)/4 = 0, \mu_2 - (\mu_3 + \mu_4 + \mu_5)/3 = 0, \mu_3 - (\mu_4 + \mu_5)/2 = 0, \mu_4 - \mu_5 = 0$. Re-express H_0 in the form $H_0 : \boldsymbol{\mu}'\mathbf{U} = \mathbf{0}'$ by finding the appropriate matrix \mathbf{U} .

We will see that this way of expressing hypotheses about the elements of a mean vector is standard for the ANOVA methods we will discuss and is used by SAS and other software for setting up hypotheses of interest.

5. Consider the dental study data of Example 1 in Chapter 1 on p. 3–4 of the notes. These data may be found on the class web page in the file `dental.dat`. The file has 5 columns:

(1) observation number, (2) child number (1–27), (3) age, (4) distance measurement, and (5) indicator of gender (0=girl, 1=boy).

Suppose that the question of interest to be addressed by an analysis of these data is whether the “rate of change” of distance over age is similar for boys and girls, as on p. 3 of the notes. The plot of the dental data on p. 3 suggests that one might consider a statistical model that says the relationship between the distance measurement in column (4) and age in column (3) is a straight line for each gender. To address this question, an analyst knowing nothing about longitudinal data analysis but knowing something about ordinary regression analysis might be tempted to postulate a usual linear regression model that says that mean distance changes with age as a straight line for each gender, where the straight lines for each gender have their own intercepts and slopes, **ignoring** the fact that the observations come in groups of four from different children. That is, s/he would write a model treating all $n = 4 \times 27 = 108$ pairs (Y_j, t_j) , $j = 1, \dots, n = 108$, as **statistically independent**, where Y_j = the j th distance measurement taken at time t_j :

$$\begin{aligned} Y_j &= \beta_{00} + \beta_{01}t_j + \epsilon_j, \text{ for gender 0,} \\ &= \beta_{10} + \beta_{11}t_j + \epsilon_j, \text{ for gender 1.} \end{aligned} \tag{1}$$

Here, β_{00} and β_{01} are the intercept and slope of the mean relationship for girls (gender 0) and β_{10} and β_{11} are the intercept and slope of the mean relationship for boys (gender 1). The ϵ_j , $j = 1, \dots, n$, are deviations with mean 0 and variance σ^2 that are assumed independent across j (the usual assumption in ordinary linear regression).

(a) Do you think that the usual assumption that the ϵ_j are independent is reasonable for the dental study? Why or why not?

(b) In terms of model (1), write down a null hypothesis that addresses the question of interest.

(c) Define $g_j = 0$ if the gender associated with observation j is 0 (girl), and $g_j = 1$ if it is 1 (boy). Verify that we can write the model in (1) equivalently as

$$Y_j = \beta_{00}(1 - g_j) + \beta_{10}g_j + \beta_{01}(1 - g_j)t_j + \beta_{11}g_jt_j + \epsilon_j. \tag{2}$$

(d) Suppose instead we write a model of the form

$$Y_j = \alpha_0 + \alpha_1(1 - g_j) + \alpha_2t_j + \alpha_3(1 - g_j)t_j + \epsilon_j. \tag{3}$$

Demonstrate that this model and model in (2) are in fact saying the same thing by finding the expressions for the α s in (3) in terms of the β s in (2).

(e) We are now going to fit both of these models to the dental data using SAS. Write a SAS program that does the following:

- Read in the data from the file using a `data` step with an `infile` statement as in the first `data` step in Example 1 of Chapter 4.
- To fit model (2), call `proc glm` with the following `class` and `model` statements:

```
class gender;
model distance = gender gender*age / noint solution;
```

- To fit model (3), call `proc glm` again as above, but now use instead the `model` statement
- ```
model distance = gender age gender*age / solution;
```

(f) From the output, identify and write down the estimates of the  $\beta$ s in model (2) and their estimated standard errors. Similarly, identify and write down the estimates of the  $\alpha$ s in model (3) and their estimated standard errors. Verify that the correspondences between the  $\beta$ s and  $\alpha$ s you found in (b) are satisfied by the estimated values.

(g) Can you give a numerical estimate and standard error from either of these models of a quantity that addresses the question of whether the “rate of change” of distance over age is similar for boys and girls? If so, write down these values. If not, explain why not.

*In later chapters, we will consider regression models for longitudinal data and different ways to write them that are similar to those above. It will be interesting to compare the analyses above to the proper longitudinal analyses we will consider. You will want to refer back to this problem when we get to this part of the course.*

6. Consider the dental data again. Suppose the question of interest is stated as asking whether there is a difference in mean response between boys and girls. To address this question, an analyst who knows nothing of longitudinal data methods might decide to postulate a statistical model that regards the 4 observations on each child as 4 “subsamples” on that child, without acknowledging that the 4 observations correspond to different ages. That is, s/he might write down the following model.

Let  $Y_{h\ell j}$  represent the  $j$ th distance measurement on the  $h$ th child from the  $\ell$ th gender,  $\ell = 0, 1$  for girls, boys. Then the model is

$$Y_{h\ell j} = \mu + \tau_\ell + e_{h\ell} + \delta_{h\ell j},$$

where  $\mu$  is an overall mean,  $\tau_\ell$  is the effect of the  $\ell$ th gender,  $e_{h\ell}$  is the “error” associated with the  $h$ th child in the  $\ell$ th gender, and  $\delta_{h\ell j}$  is the “error” associated with the  $j$ th observation on the  $h$ th child in the  $\ell$ th group.

Using SAS `proc glm`, construct the analysis of variance table for this model and calculate the  $F$  statistic appropriate for testing whether the mean responses for each gender differ. Note that you will need to be careful in identifying the correct mean square for the denominator of the error term for this  $F$  ratio!

*In an upcoming chapter, we will consider a statistical model similar to this one that acknowledges that the observations on each child are not just random subsamples but are taken at different ages.*

7. In the file `insulin.dat` on the class web page you will find longitudinal data from an experiment conducted by a diabetes researcher who was interested in comparing patterns of blood sugar reduction brought about by the use of several different insulin mixtures. The study involved  $m = 36$  rabbits, where 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 were 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).

(a) Write a SAS program to obtain the following:

- (i) Read in the data in the form they appear in the data set and, if necessary, transform the data set into a form suitable for carrying out the analyses in (ii)–(v) below.
  - (ii) Find the means for each insulin group at each day and plot them for each group on the same graph
  - (iii) Find the sample covariance and sample correlation matrix for each insulin group
  - (iv) Obtain the centered and scaled versions of the responses at each month for each insulin group (you may wish to output them to a file if you are going to use R in (b) below).
  - (v) Find the pooled sample covariance matrix and corresponding estimated correlation matrix under the assumption of a common covariance matrix
- (b) Use R, SAS `proc insight` (as in the Examples on the class web page), or some other software to obtain scatterplot matrices for each insulin group.
- (c) Based on inspection of the results in (a) and (b), do you think the assumption of a common covariance matrix for each group is reasonable? Why or why not? Give details of your reasoning.
- (d) Based on inspection of the results in (a) and (b), what covariance model(s) do you think is(are) appropriate for these data? Give details of your reasoning.