BIOSTAT 653 Homework #1

Due Wednesday September 27th, 3:10pm, in class.

It is to your advantage to type the solutions. You are free to choose any software (R or SAS or anything else) to solve these problems.

Problem 1

Consider a set of linear regression models

$$Y_i = X_i\beta_i + \epsilon_i, \epsilon_i \sim N(0, \sigma^2 I)$$

where $i = 1, \cdots, r$ represents groups. In each i'th group there are n_i samples. Each Y_i is an n_i-vector of phenotypes, X_i is an n_i by p matrix of covariates, β_i is a p-vector of corresponding coefficients. Find a test for $H_0: \beta_1 = \cdots = \beta_r$ and write down the test statistics.

Problem 2

Consider the model $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \epsilon_i$, where $i = 1, \cdots, n$ represents observations, and $\epsilon_i \sim N(0, \sigma^2)$. There are $n = 15$ observations. We can compute several key quantities: $Y^T Y = 3.03$, $X^T Y = \begin{pmatrix} 6.03 \\ 158.25 \end{pmatrix}$, $X^T X = \begin{pmatrix} 15.00 & 374.50 \\ 374.50 & 9482.75 \end{pmatrix}$.

(1) Estimate β_1, β_2 and σ^2.
(2) Give 95% confidence intervals for β_2 and $\beta_2 - \beta_1$
(3) Perform an $\alpha = 0.05$ two-sided test for $H_0: \beta_1 = 0.5$
(4) Find an appropriate p value for the one-sided test of $H_0: \beta_1 + \beta_2 = 0$ vs $H_1: \beta_1 + \beta_2 > 0$.

Problem 3

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 c1 and c2 be constants.

(1) Compute $var(c_1Y_1 + c_2Y_2)$ use the definition of variance
(2) Let $Y = (Y_1, Y_2)^T$ and $c = (c_1, c_2)^T$. Write down the covariance matrix Σ of Y, and verify that $var(c^T Y) = c^T \Sigma c$.

Problem 4

Suppose we have the following statistical model:

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

where $i = 1, \cdots, n$ represents subjects and $j = 1, \cdots, m$ represents repeated measurements. We assume $b_i \sim N(0, \sigma_b^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$, where b_i and e_{ij} are independent random variables for each I and j.

(1) Find the variance $var(Y_{ij})$ for i’th individual

(2) Find the covariance of Y_{ij} and Y_{ik}, $cov(Y_{ij}, Y_{ik})$, for i'th individual
(3) Based on results from (1) and (2), write down the covariance matrix of Y_i

Problem 5

Suppose we conduct a cross-sectional study of test scores among students in elementary and junior high schools. On one day of testing, investigators administer grade-level mathematics tests to students who are 8 years old, 10 years old, 12 years old, and 14 years old. They hypothesize that performance may depend both on age and on gender. They fit the model

$$y_i = \beta_0 + \beta_1 I(boy_i) + \beta_2 age_i + \beta_3 age_i I(boy_i) + \epsilon_i$$

where $I(boy_i)$ is an indicator function that equals to 1 when i'th individual is a boy, and equals to 0 otherwise. Describe in words the hypothesis tested by each choice of L and θ_0 below.

1) $L = (0\ 0\ 0\ 1), \theta_0 = 0$

2) $L = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \theta_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

3) $L = (1\ 1\ 10\ 10), \theta_0 = 80$

4) $L = (0\ 0\ 4\ 4), \theta_0 = 5$

5) $L = (0\ 0\ 4\ 0), \theta_0 = -5$

6) $L = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \theta_0 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$

Problem 6

It is well-known that lead exposure may have a negative effect on IQ. However, it is not known if the effect of lead exposure is persistent and irreversible. To answer this question, investigators studied a group of children who lived near a lead smelter. Based on the blood-lead measurement, these children can be classified into three categories: unexposed (=1), currently exposed (=2), and previously exposed (=3). The investigators collected important information from these children (gender and age) and performed test to determine the IQ for each child.

You can find the data file (leadiq.txt) on canvas. Each row represents: ID, lead exposure category, gender (boy=0, girl=1), age (in years) and IQ.

The investigators wish to use a linear regression model to answer the following research questions:

a) Research Question 1: Is there an effect of lead exposure on IQ?
b) Research Question 2: Does the effect of lead exposure on IQ depend on gender?
c) Research Question 3: Does the effect of lead exposure on IQ decay with time?

Your task is to help the investigators answer these research questions. To do so,

1) Describe, justify and fit a linear regression model that can be used to address all these research questions.
2) Do we need to use age as a covariate? Why or why not?
3) Provide parameter estimates and interpret the results in language someone without a statistics background can understand.
4) What is the expected IQ for a boy without lead exposure? What is the expected IQ for a boy who is currently exposed with lead? What is the expected IQ for a boy who had lead exposure before?
5) Provide evidence that a linear regression model does properly fit the data.

Problem 7

As we know from problem 2, lead exposure is dangerous for children. To help these children who lived near a lead smelter, the investigators decided to find out treatments that could reduce the blood-lead levels. One particular trial they performed was a placebo-controlled, randomized study of succimer (a chelating agent). They performed this study in children with relatively high blood lead levels. They collected measurements of blood lead levels in 100 children at four different time points: week 0 (a.k.a. baseline), week 1, week 4, and week 6. These 100 children were randomly assigned to chelation treatment with succimer or to placebo. For simplicity, however, we will focus only on the 50 children assigned to chelation treatment with succimer.

You can find the data file (lead.txt) for these 50 samples on canvas. Each row represents: ID, blood lead levels at week 0, blood lead levels at week 1, blood lead levels at week 4, blood lead levels at week 6.

1) Calculate the sample mean, standard deviation and variance of the blood lead levels at each occasion (i.e. time point).
2) Construct a time plot of the blood lead levels for all individuals over time. Construct a time plot of the mean blood lead level over time. Describe the general characteristics of the time trend.
3) Calculate the 4 by 4 covariance and correlation matrices for the four repeated measures of blood lead levels. Are the diagonal elements in the covariance matrix identical to the variance computed from (1)?