LONGITUDINAL DATA ANALYSIS: FINAL EXAM, MARCH 11th 2005

This homework assignment is to be considered a **take-home** test. Thus, you **may not** collaborate with any other person (whether in the class or not), **nor** may you consult with *anyone*. You **may** use any reading material (class notes, books, etc) you wish. This assignment is due **March 21st at 5:00 pm**

1. A study was conduced in West Java, Indonesia, to determine the effects of vitamin A deficiency in preschool children. The investigators were particularly interested in whether children with vitamin A deficiency were at increased risk of developing respiratory infection, which is one of the leading causes of death in this part of the world.

   250 children were recruited in the study, and their age in years, gender (0 =male, 1 =female), and whether they suffered vitamin A deficiency (0 =no, 1 =yes) was recorded at an initial clinic visit (time 0). Also recorded was the response, whether the child was suffering from a respiratory infection (0 =no, 1 =yes). The children then were examined again at 3 month intervals for a year (at 3,6,12, and 15 months after the first visit) and the presence or absence of respiratory infection was recorded at each of these visits. Luckily, all children we seen at all visits, so there were no missing data.

   The data file

   `http://www.biostat.jhsph.edu/~fdominic/teaching/LDA/ICHS.dat`

   has the following columns:

Column	Description
1	Child id
2	Response (0 or 1 as above)
3	Time (in months, as above)
4	Gender (male = 0, female=1)
5	Vitamin A (not deficient =0, deficient =1)
6	age (in years)

(a) Let $\boldsymbol{y}_i$ be the vector of responses for the ith child, consisting of elements y_{ij}, the observations on whether the child has a respiratory infection at time t_{ij} (recorded in months). Write down a model for $E(y_{ij})$ in terms of an appropriate link function that is linear in an intercept and include additive terms for time, age, gender, and vitamin A status. Also, write down $var(y_{ij})$ given the nature of the response.

(b) Under your model for $E[y_{ij}]$ in (a):

(i) What is the probability that a female child age 4 who does not have vitamin A deficiency will not have a respiratory infection at the final visit? (*Hint: give answers in terms of model parameters*)

(ii) what are the *odds* that a male child of age 3 with vitamin deficiency will have a respiratory infection at the initial visit? (*Hint: give answers in terms of model parameters*)

(iii) What must be true if the probability of having respiratory infection is greater for children with vitamin A deficiency than for children without for any age/gender/time? (*Hint: give answers in terms of model parameters*)

(c) The investigators had not taken a course in longitudinal analysis; thus, they were unaware that measurements on the same child might be correlated. They fit the model in (a) without taking correlation into account, treating all the observations from all children as if they were *unrelated*.

Based on this fit, is there sufficient evidence to suggest that the mean pattern of respiratory response is associated with the presence or absence of vitamin A deficiency? State the null hypothesis corresponding to this issue in terms of your model (a), cite the test statistic and p-value on which you base this conclusion, and state your conclusion as a meaningful sentence.

(d) One of the investigators then talked to a friend who knew something about repeated measurements, who suggested that the analysis in (c) may be unreliable because possible correlation had not been taken into account. Give a brief explanation of why failure to take correlation into account might be expected to lead to unreliable hypothesis tests.

(e) Because you have taken a course in longitudinal data analysis, the investigators called you in for help with an improved analysis. Extend the model (a) to take into account correlation among repeated measurements on the same subject.

(f) Fit your model in (e) to the data, *making as few assumptions as you can* about the possible structure of correlation among the elements of a data vector. Assuming that your assumed model for correlation is correct, conduct a test of null hypothesis in part (c), citing an appropriate test statistic and p-value. State your conclusion as a meaningful sentence.

Do the results agree with those in part (c)? Give a possible explanation for this, citing results from your output to support your explanation.

(g) From inspection of your fit in (f), do you think a simpler model for correlation may be plausible? Select a correlation model you feel is most plausible based on your inspection, explaining why you chose this model, and fit this model to the data.

(i) Is there sufficient evidence to suggest that the probability of respiratory infection changed over the 15 month study period?

(ii) Is there sufficient evidence to suggest that it is worthwhile to take gender into account in understanding the risk of respiratory infection in this population of children?

(h) From your fit in (g), provide an estimate of the probability that a female child of age 7 with vitamin A deficiency has a respiratory infection at the initial visit.

Given these considerations, conduct an analysis of these data. Write a brief report summarizing:

- The statistical model you assumed, and why you choose it
- The analyses you conducted, the assumptions you made and why you made them
- The results, addressing the interests of the investigators as described above.

Carry out whatever analyses you feel are appropriate. It is important to wrote a clear and organized report summarizing what you did and why you did it. DO NOT INCLUDE COMPUTER OUTPUT