LONGITUDINAL DATA ANALYSIS
Homework II, 2005

A public health study was conducted to estimate the association between maternal smoking and respiratory health of children in two cities Kingston and Portage. Each child was examined once a year at a clinic visit (visits at ages 9, 10, 11, and 12) for evidence of "wheezing". The response was recorded as a binary variable (0 = wheezing absent, 1=wheezing present). In addition, the mother's current smoking status was recorded (0=none, 1=moderate, 2=heavy). The scentific question is to assess and compare the effects of smoking patterns on wheezing patterns.
The data file `wheeze2.raw` is posted on the course web site and has the following columns:

Columns	Description		
1	child id		
2	city		
3-5	age = 9	smoking indicator	wheezing response
6-8	age = 10	smoking indicator	wheezing response
9-11	age = 11	smoking indicator	wheezing response
12-14	age = 12	smoking indicator	wheezing response

Let y_{ij} be the wheezing indicator on the ith child at the jth age t_{ij}, where t_{ij} ideally takes on all values $9, 10, 11, 12$. For each child i, let

$$\begin{aligned} x_{0ij} &= 1 \text{ if smoking = none at } t_{ij} \\ x_{0ij} &= 0 \text{ otherwise} \\ x_{1ij} &= 1 \text{ if smoking = moderate at } t_{ij} \\ x_{1ij} &= 0 \text{ otherwise} \\ c_i &= 0 \text{ if city = Portage} \\ c_i &= 1 \text{ if city = Kingston} \end{aligned}$$

(a) 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 city, for smoking (none and moderate), and time. Also, write down $var(y_{ij})$ given the nature of the response.

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

(b.1) What is the log-odds of wheezing for a child from Portage whose mother is heavy smoker at t_{ij}?

(b.2) What must be true if the probability of wheezing is smaller for a child from Kingston rather than Portage? (*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 wheezing is associated with mother's smoking? 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 model for correlation is correct, conduct a test of null hypothesis in part (c). 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) Do you think a simpler model for correlation may be plausible? Select a correlation model you feel is most plausible, explaining why you chose this model, and fit this model to the data.

(g.1) Is there sufficient evidence to suggest that the probability of wheezing is associated with maternal smoking?

(g.2) Is there sufficient evidence to suggest that it is worthwhile to take city into account in understanding the risk of respiratory wheezing in this population of children?

(h) From your fit in (g), provide an estimate of the probability that child from Kingston whose mother is heavy smoker wheeze at the initial visit and an estimate of the probability that child from Kingston whose mother does not smoke wheeze at the initial visit. What can you conclude?

(i) One could imagine that wheezing at a particular time might be dependent on past and present maternal smoking behavior. Alternatively, One could imagine that wheezing at a particular time might be dependent on previous wheezing. Perhaps children who have already exhibited such behavior are more prone to show it again. Fit two logistic regression models which allow to investigate these two phenomena. Report and interpret the odds ratio estimates of wheezing.

(l) Specify a logistic regression model with random intercept and additive terms for city, for smoking (none and moderate), and time.

(l.1) What is the log-odds of wheezing for a child with random intercept $U_i = 0$, from Portage whose mother is heavy smoker at t_{ij}?

(l.2) What is the log-odds of wheezing for a child with random intercept $U_i = 2$ from Portage whose mother is moderate smoker at t_{ij}?
(*Hint: give answers in terms of model parameters*)

(m) Fit the logistic regression model with random intercept, estimate (l.1) and (l.2) and compare these estimates with the population average estimates obtained from model (g). Report and interpret the estimated degree of heterogeneity across children in the propensity of wheezing not attributable to the covariates.

(n) Given all the data analyses you have conducted so far, 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. Two pages maximum