

LDA 140.655: FINAL EXAM 2009

This midterm is assigned March 4th, 2009.

Your solution write-up and appendix is due on March 11th, 2009 at 5PM sharp.

Two options for turning in:

1) A hard copy of the write-up and appendix must be placed in Howard Chang's departmental mailbox on the 3rd floor outside of the Biostatistics Departmental office.

2) Turn in a hard copy in class to Francesca on March 11th

Late submissions are not allowed.

Electronic (email) submissions are not allowed.

Final Rules:

This exam is to be considered a take-home test. Please do not collaborate or consult with others. You may use any reading material (class notes, books, etc) you wish. Please keep your write-ups short and do not include reams of analyses output. Please attach an appendix of the (well documented) code you used.

For the final exam, the BBS will be used only for **clarification** of the exam questions, wording or expectations.

You may also carry out additional analyses that you feel are appropriate.

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 (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 scientific question is to assess and compare the effects of smoking patterns on wheezing patterns.

The data file `wheeze2.raw` (wide format) 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 of the i^{th} child at the j^{th} age. For each child i , let

$$\begin{aligned}x_{1ij} &= 1 \text{ if smoking = moderate at } t_{ij}; x_{1ij} = 0 \text{ otherwise} \\x_{2ij} &= 1 \text{ if smoking = heavy at } t_{ij}; x_{2ij} = 0 \text{ otherwise} \\t_{ij} &= \text{age (9, 10, 11, 12)} \\c_{ij} &= \text{city (1 = Kingston, 0 = Portage)}\end{aligned}$$

(a) Using the above variables, write down a model for $E(y_{ij})$ with an appropriate link function that contains covariates including an intercept and additive terms for city, maternal smoking status (moderate and heavy), and time. Also, write down $\text{Var}(y_{ij})$ in terms of $E(y_{ij})$ given the nature of the response. Interpret all coefficients in your model.

(b) Under your model for $E(y_{ij})$ in (a), what describes the log-odds of wheezing for a child from Kingston whose mother is a heavy smoker at t_{ij} ? (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 the null hypothesis, cite the test statistic and p-value on which you base this conclusion, and state your conclusion in a meaningful sentence.

(d) One of the investigators then talked to a friend who knew something about repeated measurements and 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. Write down an extended population-average model to (a) that takes into account correlation among repeated measurements on the same subject. *Make as few assumptions as you can* about the possible structure of correlation among the measurements within a child.

(f) Fit your model in (e) to the data and conduct a test of the 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 and explain why you chose this model. 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 estimates for (1) the probability that a child from Kingston whose mother is heavy smoker wheezes at the initial visit and (2) the probability that a child from Kingston whose mother does not smoke wheezes at the initial visit. What can you conclude about the effect of mother being a heavy smoker at the initial visit on wheezing?

(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 you to investigate these two phenomena. Report and interpret the odds ratio estimates for wheezing.

(j) Specify a logistic regression model with random intercept and additive terms for city, for smoking (moderate and heavy), and time. Interpret all model parameters.

(j.1) What describes the probability of wheezing for a typical child (with random intercept deviation $U_i = 0$), from Portage whose mother is heavy smoker at t_{ij} ? (*Give answers in terms of model parameters.*)

(j.2) What describes the probability of wheezing for a child with higher baseline probability of wheezing (random intercept deviation $U_i = 2$), from Portage whose mother is moderate smoker at t_{ij} ? (*Give answers in terms of model parameters.*)

(j.3) What describes the odds ratio of wheezing for a child from Kingston whose mother is heavy smoker versus if the child is from Portage with a mother who doesn't smoke at t_{ij} ? (*Give answers in terms of model parameters.*)

(k) Fit the logistic regression model with random intercept and compare the estimated coefficients and their standard errors with those obtained from model (g). Are the two models equivalent? Also estimate (j.1) and compare the estimate with the population average estimate obtained from model (g). Finally, report and interpret the estimated degree of heterogeneity across children in the log-odds of wheezing not attributable to the covariates.

(l) Given all the data analyses you have conducted so far, write a brief summary discussing:

1. The analyses you conducted, the assumptions you made, and why you made them
2. The results, addressing the interests of the investigators as described above.