

LDA 140.655

Midterm 2009

This midterm is assigned February 16th, 2009.

Your solution write-up and appendix is due on February 23rd, 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 Feb 23rd.

Late submissions are not allowed.

Electronic (email) submissions are not allowed.

Midterm rules:

- 1) You are allowed to consult notes and the textbook*
- 2) You are allowed to consult a LDA classmate*
- 3) Your write-up must be independent work and not team work*
- 4) As an appendix, well-documented code will accompany the write-up*

A study was conducted to investigate two treatments for male patients suffering from multiple sclerosis. 150 male suffers of the disease were recruited into the study, and 75 were randomized to receive azathioprine (AZ) alone (group 1), and 75 were randomized to receive azathioprine + methylprenisommne (AZ + MP, group 2). For each participant, a measure of autoimmunity, AFCR, was planned at clinic visits at baseline (time 0, at initiation of the treatment) and at 3,6,9,12,15, and 18 months thereafter. Multiple sclerosis affects the immune system: low values of AFCR (approaching 0) are evidence that immunity is improving, which is hopefully associated with a better prognosis for suffers of MS. Also recorded for each subject was age at entry into the study and an indicator of whether or not the subject had had previous treatment with either of the study agents (0=no, 1=yes). The average age of the men across groups was 50.45, with SD 6.69.

The primary scientific aim of the study are to investigate whether:

- both treatments (AZ or AZ + MP) lower AFCR over the 18 months period;
- treatment with AZ + MP results in different immune system response than does AZ alone, and if so how it is different in terms of response over time.

It was also suspected that a subject's age and prior history might be related to their AFCR level at baseline and to the rate at which AFCR changes during the 18 months period. The square root of ACFR is the response variable of interest (square roots were taken so

that the AFCR observations better satisfy the assumption of normality – note that this is sub-optimal but we’ll work with it for now).

The data are in the file `aocr.raw`, which you can download from the class data web page. In the file, each record corresponds to a single observation, with columns:

- col1 = subject id
- col2 = time (months)
- col3 = square root AFCR
- col4 = group (1 = AZ alone, 2 = AZ + MP)
- col5 = prior treatment indicator
- col6 = age (years)

Download these data from the course website and read them into Stata using the `infile` command (See Lab 1 documents for examples of using the `infile` command). Make sure you have the data in “long” format.

Part I: Exploratory Data Analysis (EDA)

- a. Describe the structure of the data set, including distribution of observation times and the number of subjects observed at each time point. Understand the distributions of covariates and whether or not they are time varying.
- b. Visualize the longitudinal trajectories of the subjects’ square root of AFCR over time. Include a smooth lowess curve on each of the following three plots to show the overall trend in the square root of AFCR.
 - i. Make a spaghetti plot of the square root of AFCR using information on all the subjects
 - ii. Make a spaghetti plot of the square root of AFCR using information on a random sample of 10% of the subjects
 - iii. Make a ZAP plot where you plot the trajectories for a subset of individuals selected based on quantiles of the individual-specific median square root of AFCR.
- c. Explore overall trends in the data with respect to the primary scientific aim of the study. Present a plot that illustrates the trend in the response (square root of AFCR) over time with respect to treatment group. Ignore age and prior treatment effects for now. Write a **few** sentences summarizing the results.
- d. Explore the correlation structure of the response variable using correlation matrices and the sample autocorrelation function (ACF). Make sure to remove covariates effects as you see appropriate. Include a confidence interval around the ACF estimate. Describe your results.

- e. Create and plot a variogram of the afcr data. Use this variogram to create an ACF plot. Plot the variogram-estimated ACF on the same graph as the ACF estimated above. (Use the same residuals that you used to create the ACF in the previous problem.)

Part II: Modeling the AFCR data

Use the AFCR dataset to assess and compare the effects of two treatments for patients suffering from MS.

- (a) Formulate a general mean model that includes, at minimum, an effect for treatment group, for age, prior treatment, for time, and all two-way interactions between time and the other three covariates (formulate a model that assumes independence – i.e., ignore the correlation in the responses). Write down your model. Run the model.
- (b) Using residuals from the model in (a), study the correlation structure, and select a correlation model under which to analyze the data. Justify your choice.
- (c) Using GEE and your selected correlation structure from (b), test whether treatment has an effect on AFCR, either at baseline and/or on the rate of change of AFCR over time. Use the Huber-White (sandwich) method of robust variance estimation to construct your test. Clearly state null and alternative hypotheses in terms of the model parameters in (a).

Hint: Run a GEE population-averaged model as formulated in (a) with correlation structure (b). See the help file for xtgee to see what option computes the Huber-White variance. Then, use the command “test” on the appropriate coefficients corresponding to each of the following hypothesis tests: “The inclusion of treatment group in the model is justified because treatment group explains a significant amount of variation in sqafcr,” and more specifically, “treatment group affects the rate of change of sqafcr over time.” This will involve two instances of the “test” command. The first instance will involve all coefficients of any variables involving treatment group. The second instance will only involve the coefficient of an interaction term.

- (d) Repeat (c) for the prior treatment variable instead of the treatment variable in the study.

Hint: Re-read the hint for (c) above. You’ve already run the model, so no need to rerun it. Do the same test commands procedure but instead of the treatment group variable, apply the logic to the prior treatment variable. Be sure to clearly state for each test command the null and alternative hypotheses that are being tested in terms of the model parameters as specified in (a).

- (e) Based on your results in (c) and (d), fit a reduced model, if any, and present key parameter estimates and CIs based on your GEE model fit. Interpret the effects of treatment group and prior treatment on AFCR at baseline and over time. Again, use the robust variance estimator. Make sure your interpretation of parameter estimates is for “population average” or “marginal model” effects.

Hint: Look at the results from the instances of test commands in (c) and (d). If a certain coefficient was determined to not significantly differ from 0, then it is effectively 0 in terms of the model we are building, which is the same as removing the corresponding variable from the model. If a certain coefficient was found to be not significantly different from zero, remove it and the corresponding variable from the model specified in (a), and write the remaining coefficients and variables down for you answer in (e).

- (f) Refit your model from (e) using:
- your chosen correlation structure with model-based variance estimates;
 - independence correlation structure with model-based variance estimates;
 - independence correlation structure with robust variance estimates.
- Compare the standard errors generated by each of the four model fits and comment on similarities and differences (i.e. try to explain them).

Hint: This results in four models:

Model 1: the model from (e)

Model 2: the model from (e) without the Huber-White option

Model 3: the model from (e) with an independence correlation structure and without Huber-White

Model 4: the model from (e) with an independence correlation structure and with Huber-White

Natural comparisons would include: Model 1) to Model 2) as well as Model 3) to Model 4) to see how the Huber-White changes estimates, and then Model 1) to Model 4) as well as Model 2) to Model 3) to see how a different correlation structure affects model estimates while holding constant the Huber-White inclusion/omission.

Part III: Theoretical considerations

Required for Biostatistics PhD, ScM, and MHS Students

- a) Derive the variance of the β_{WLS} estimate using matrix notation for a general weight matrix W .
- b) Derive the variance of the β_{WLS} estimate using matrix notation when the weight matrix W is $V^{-1} = \text{Var}(Y)^{-1}$.