

## Assigned Data Set Project

Due: Thursday, Mar 6, 2003.

Please keep your report to about 3 typed double spaced pages plus a few pictures and tables.

Data is available at <http://rem.ph.ucla.edu/~rob/rm/examples/index.html> for the link and some other files. The actual sas data file is <http://rem.ph.ucla.edu/~rob/rm/examples/cd4m236.sas7bdat>.

This data set contains cd4 cell counts for a group of men who eventually became seropositive, meaning that they became infected with the HIV virus some time during the study. Seroconversion means that they convert from being HIV negative to HIV positive. The seroconversion point is at months=0 or binmonth = 0. The average CD4 count should generally be flat while healthy with an average value of around 1100. After seroconversion (HIV infection), it should decrease.

Your goal is to develop a good model or models for the data that can help you answer the following two questions.

1. First, how does the population average change after seroconversion? Please answer both qualitatively and quantitatively and back up your results with appropriately chosen model or models.
2. Second, choose one (only!) of the covariates, and figure out how that affects CD4 level. I recommend seeing how being a smoker affects the CD4 level. An alternative choice is cesd – there is a hypothesis that higher levels of depression are associated with worse immune system function, that is, with lower CD4 cell counts.

Some instructions

1. You will model the response as continuous.
2. Use binmonth as your time variable – it gives us a balanced with missing data structure. (There are several other versions of time in the data set).
3. Only consider the effects of a single covariate in addition to time. You do not need to consider the other covariates that you choose not to make use of in your analyses.

4. If you would like a challenge, the original version of the data (slightly larger, in tab delimited format) is also available on the web site. Similarly, if you would like to explore additional covariates, you may, but it is not required. If you decide to do this, please say so up front in your analysis and be sure to do a good job.
5. Be sure to clearly state what the goals of your analysis are. Unless you tell me what predictor you are deciding to care about, I will not know.
6. Make sure tables and plots are well labeled and designed.
7. You must refer to all tables and plots in the text.
8. Be sure to define all entries in your tables and describe all figures carefully and accurately.
9. Explain what specific conclusions you draw from your tables and figures. Do not say something like “Here are the coefficients and standard errors from fitting the model.” Rather, state something like: Having A higher makes the response higher, that B has no significant effect on response, and that C enters with a quadratic effect that is approximately level for  $C < .6$  but the response rises rapidly for  $C > .6$ .

Variables in the data set are

Id Subject identifier

binmonth This is the time since seroconversion, rounded into six month bins. This is what you should use for time in this analysis; if you use this as your time variable, then the data is balanced with missing data. Binmonth is set equal to the midpoint of the 6 month interval. So if  $\text{Binmonth} = 3$ , then the observation was taken between month 0 and month 6.

cd4 cd4 cell counts a count of a particular type of cell in the blood system. Typical value is approximately 1100. After infection it should decrease.

conversion Indicates whether the subject has seroconverted. Basically  $\text{conversion} = 1$  if  $\text{binmonth} > 0$ , otherwise  $\text{conversion} = 0$ . If  $\text{binmonth} > 0$ , conversion should be equal to one.

smoker no if non-smoker, yes if smoker. Time fixed covariate created from the packs variable.

cesd is a depressive symptoms score. Higher means more depressed. Timevarying.

Additional variables available in the data set that you need not necessarily attend to.

meancesd average of cesd scores. Time fixed.

packs 0 if non-smoker, else # of packs of cigarettes smoked

druguser Time fixed: 1 if subject ever used recreational drugs, else 0.

drugs 1 if the subject used recreational drugs during the last unit of time, 0 if not.

age age of individual in years from an arbitrary zero point. (This arbitrary re-zeroing, which should be the same for everyone, is to help make it difficult to identify individuals in the study.) Appears to be age at study entry.

log2cd4 Log base 2 of the cd4 count

sexpart # of sex partners

time time since seroconversion measured in years

timedays time since seroconversion measured in days

Actual variables and names and labels in the current data set.

-----Variables Ordered by Position-----

#	Variable	Type	Label
1	id	Char	subject id
2	binmonth	Num	binmonth +/- 3
3	conversion	Char	before/after seroconversion
4	cd4	Num	CD4 count
5	log2cd4	Num	log <sub>2</sub> (cd4)
6	age	Num	age(yrs) - a constant
7	smoker	Char	(time fixed) smoker (yes/no)
8	druguser	Char	(time fixed) druguser (yes/no)
9	meancesd	Num	mean depression score
10	packs	Num	packs of cigs per day
11	drugs	Num	drug use (yes/no)

12	sexpart	Num	number of sex partners
13	cesd	Num	depression score - a constant
14	time	Num	years before/after seroconversion
15	timedays	Num	days before/after seroconversion

## Final Project

Final Project is due Tuesday, 3pm, Mar 18 in my mailbox in the Biostat office 51-254. Please do not slip projects under my door as that is a good way to loose them.

Projects should be a maximum of approximately 5 typed double spaced pages plus a few plots and tables. Same suggestions apply as given for the assigned data analysis project regarding plots and tables. Be clear about the purpose of your analysis, and about the conclusions that you draw from your analysis.