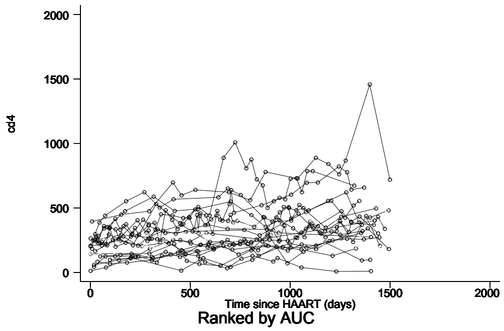

Longitudinal Data

Spring 2013

January 24

Chapter 1

Notation

Instructors

Alan Hubbard

(hubbard@berkeley.edu)

Reuben Thomas

GSI

Katia Eliseeva

Why have so much notation?

- So many things (random variables, random vectors, regression coefficients, variance and co-variance parameters,...) so little time.

- If one could explain a statistical model, estimation procedure, etc. simply and efficiently in English every time, we would not need notation.

- Every little detail of notation has meaning (telling you some part of the statistical story)

Notation, cont.

- However, we have complicated data and models and we need a shared language that efficiently translates what we're talking about.

- One goal of this course is to translate a scientific hypothesis regarding longitudinal data into a specific statistical model that yields the parameters of interest. This starts with notation.

Typical Symbols Commonly Used for Different types of Objects

- Parameters
- Random Variables, Random Vectors, Random Matrices
  - Measured Variables
  - Latent Variables
- Constants
- Operations (expectations, sums, logs, etc.)

Random Variable

- Say, one collects an outcome, Y, and a covariate, X on a randomly sampled set of n subjects out of a much larger target population.

- We could represent this experiment as random draws of $O=(Y,X)$.

- We could more explicitly say we have for each individual, i, $O_i=(Y_i,X_i)$, $i=1,..n$.

- We could talk about the mean of the outcome in certain sub-groups defined by a specific value of the covariate, X, or $E(Y|X=x)$.

- We could talk about defining a function that converts the random variable X into the mean given that X, or $E(Y|X)$ which is a random function of X.

Latent Variables, Parameters

- Simplest case is the normal linear model, or
  - $O=(Y,X)$, i.i.d. (independent and identically distributed)
  - Statistical Model for Y is:

$$Y=\beta_0+\beta_1 X+\varepsilon,\ \varepsilon\ i.i.d\ N(0,\sigma)$$

- ε is a latent variable that helps to define the model.
- What are the random variables?
- What are the parameters?
- What does the equation imply about the distribution of Y?

Vectors (of Random Variables, Parameters)

- It's convenient to represent longitudinal data not just as single numbers, but also vectors and matrices.

- A random vector is a set of random variables (e.g., the set of cholesterol measurements made on a subject).

- Another random vector could be the covariates (explanatory variables) measured on the subject at a single time (e.g., age, race, weight, ...).

- A parameter vector is a set of parameters (e.g., the set of predicted mean cholesterol for each of the measurement times).

Matrices

- For this course, matrices are often a convenient way to display both data and parameters.

- An example of representing data is the matrix being simply a set of vectors (e.g., each row of the matrix is a different observation).

- Set of measurements of explanatory variables made on a subject longitudinally can be represented as a matrix where every row contains all the measurements made on that subject at a specific time point– much like spreadsheet.

- Also, matrices are a convenient way to display certain sets of parameters, such as the set of all correlations of outcomes measured on the same subject.

Typical rules regarding random variables vectors, etc.

- Because we have more flexibility in written documents than on the blackboard, we need different rules.

- In documents
  – Random Variables capitalized, realizations small, e.g., $P(Y=y)$.
  – Vectors in bold: $P(\mathbf{Y}=\mathbf{y})=P(Y_1=y_1,Y_2=y_2,\ldots)$
  – Matrices in capital or bold (need the context to tell the difference).

Typical rules regarding random variables vectors, etc.

- On Board
  - Random Variables capitalized, realizations small, e.g., *P(Y=y)* (same)
  - Vectors with arrow over top:

$$P(\vec{Y} = \vec{y}) = P(Y_1 = y_1, Y_2 = y_2, \ldots)$$

  - Matrices underlined or more likely in context.

Outcomes and Explanatory variables

- Y_{ij} will represent a response variable, the jth measurement of unit i.

- $$\mathbf{X}_{ij}^T = (1, X_{ij1}, X_{ij2}, \ldots, X_{ijp})$$ or: $$\mathbf{X}_{ij} = \begin{pmatrix} 1 \\ X_{ij1} \\ X_{ij2} \\ X_{ij3} \\ \ldots \end{pmatrix}$$

  Will (typically) be a vector of length p+1 of explanatory variables observed at the jth measurement (note, the 1 is included to allow for an intercept) – the superscript T means transpose (so untransposed is a column vector).

Numbers of observations on one individual and number of individuals

- $j = 1, n_i$. $i=1,m$ - so the number of longitudinal observations for person i is n_i, number of subjects is m.

$$\mathbf{X}_{ij} = \begin{pmatrix} 1 \\ X_{ij1} \\ X_{ij2} \\ \ldots \\ X_{in_i p} \end{pmatrix}$$

Parameters of Interest

- We will discuss estimates of parameters related to means (like regression coefficients) and those related to variances and covariances.

- For example, $E(Y_{ij})$ *or* $E(Y_{ij}|\mathbf{X}_{ij})=\mu_{ij}$, $Var(Y_{ij})$ or $Var(Y_{ij}|\mathbf{X}_{ij})=v_{ijj}$

Nesting Observations (measurement within individual)

- Set of repeated measures for unit i are collected into a ni-vector $\mathbf{Y}_i^T = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})$.
- $\boldsymbol{Y}_i$ has mean, $E(\boldsymbol{Y}_i)=\boldsymbol{\mu}_i$ and $n_i \times n_i$ covariance matrix $Var(\boldsymbol{Y}_i)=V_i$.
- The jk element of V_i is the covariance between Y_{ij} and Y_{ik}, that is $cov(Y_{ij}, Y_{ik})=v_{ijk}$.

Nesting Observations (measurement within individual)

- Note, that $cov(Y_{ij}, Y_{ij}) = var(Y_{ij}) = v_{ijj}$
- To represent how observations co-vary on a subject, we will sometimes use correlation: R_i will be the n_i x n_i correlation matrix of $\boldsymbol{Y}_i$.

Combining all observations into a big data set.

- We will lump the responses of all units into one big vector $\boldsymbol{Y}=(\boldsymbol{Y}_1,\ldots\boldsymbol{Y}_m)$ which is an N-vector (total number of observations):

$$N = \sum_{i=1}^{m} n_i$$

- Most of the course will focus on regression models of the sort:

$$\begin{aligned} Y_{ij} &= B_0 + B_1 X_{ij1} + \ldots + B_p X_{ijp} + e_{ij} \\ &= \mathbf{X}_{ij}^T \beta + e_{ij} \end{aligned}$$

Combining, cont.

- We can write the model for the data on the ith person as

$$\underset{n_i x1}{\mathbf{Y}_i} = \underset{n_i x(p+1)}{X_i} \underset{(p+1)x1}{\boldsymbol{\beta}} + \underset{n_i +1}{\mathbf{e}_i}$$

- and for the entire data as:

$$\underset{Nx1}{\mathbf{Y}} = \underset{N x(p+1)}{X} \underset{(p+1)x1}{\boldsymbol{\beta}} + \underset{Nx1}{\mathbf{e}}$$

Example: Sex and drug/alcohol use

i	X_{ij1}	X_{ij2}	Y_{ij}
ID	date	Sex	Drug/Alch
10123	3-Nov-98	no	no
10123	4-Nov-98	no	no
10123	5-Nov-98	no	no
10123	6-Nov-98	no	no
10123	7-Nov-98	no	no
10123	8-Nov-98	no	no
10123	9-Nov-98	no	no
10123	10-Nov-98	no	no
10123	11-Nov-98	no	no
10123	12-Nov-98	no	no
10125	7-Oct-98	no	no
10125	8-Oct-98	yes	no
10125	9-Oct-98	no	yes
10125	10-Oct-98	yes	no
10125	11-Oct-98	yes	no
10125	12-Oct-98	no	no