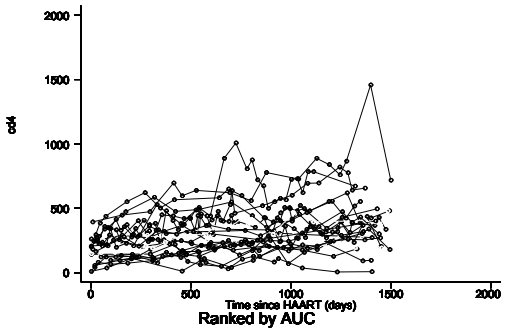


Longitudinal Data

Spring 2013

January 22

Chapter 1

Overview of Semester

Instructors

Alan Hubbard

(hubbard@berkeley.edu)

Reuben Thomas

GSI

Katia Eliseeva

Texts

Relevant Texts

No required text. Nick Jewell and Alan are currently writing a book on the subject and drafts of chapters will be posted on the website. Lecture notes will also be posted on the website.

Ebooks on Bspace Site

Course Description

- This course covers the statistical issues surrounding estimation of effects using data on subjects followed through time. The course emphasizes a regression model approach to disease incidence, continuous, binary and count outcome data.
- Background expected in statistical/mathematical material including regression, basic understanding of statistical estimation and inference.

Assignments/Exams

- We will have approximately 6-7 assignments, one midterm, one final project.
- Most assignments will involve computer analysis of data. Although the student can use any software they find convenient, STATA will be emphasized.
- The final assignment will be a data analysis (or advanced methodology "research") of the student's choosing, to be presented as a poster session in the last day of class.

List of Topics

- Introduction to course, examples of data, notation (CHAPTER 1: 1-2 weeks).
- Major themes in course including advantages and complications of longitudinal data (part of CHAPTER 3: 1 week).
- Graphical representation of longitudinal data (CHAPTER 2: 2 weeks).
- Ordinary linear regression with repeated measures, longitudinal data (CHAPTER 3: 1-2 weeks).

Topics, cont.

- Transforming longitudinal structure into cross-sectional data: baseline covariates only with repeated measures outcomes (CHAPTER 4: 1-2 weeks).
  - Summarizing a function of outcomes over time in general.
  - Specific simple examples, such as repeated events as counts – Poisson and Neg. Binomial regression.

- Possible time-dependent covariates (repeated measures regressions)
  - Contrasting different approaches: estimating equation (marginal), transitional and likelihood-based (mixed) models (CHAPTER 5: 1 week).
  - Marginal and transitional estimation (CHAPTER 6: 2 weeks).
  - Mixed models (CHAPTER 7: 2 weeks).
  - Discussion of modeling the “whole” process (mixed models) and targeting estimates of association (related comparisons of maximum likelihood estimating vs. estimating equation approaches).

Potential Topics

- Survival Analysis: Right-censored data, K-M curves, Cox regression.

- Semi-parametric methods/Graphical models
  - Definition of causal graphs and how they are used to motivate parameters of interventions.
  - Plug-in estimation of these parameters.
  - Machine learning algorithms (semi-parametric approaches)

- Ecological time series

- Trajectory analysis

Longitudinal Data - Type of data/ studies that are relevant to this course

- Data collected (or inferred) at different time on a unit (e.g., person).
- Focus typically on mean changes of the outcome variable and what variables "explains" such changes.
- Outcome can be binary (disease yes/no), continuous (CD4 levels in an HIV-infected subject), or counts (number of diarrhea episodes in a time block).
- Usually involves multiple observations on each subject.

Repeated Measures

- Continuous outcome data: we will looking an models/methods that are extensions to simple linear regression models.

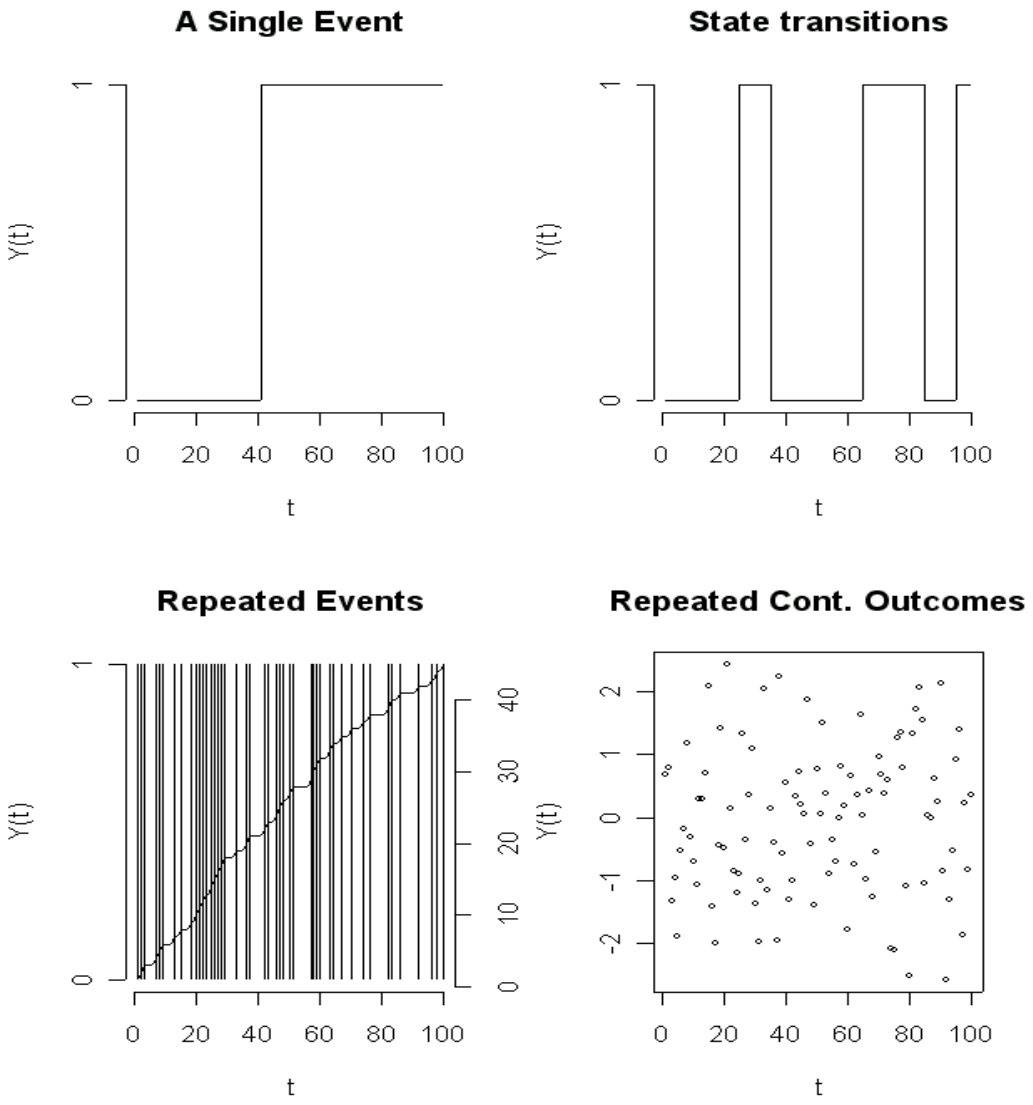
- Binary and count data using extensions of logistic and Poisson (log-linear) regression.

Longitudinal Data Reduced to Single Outcome

- Single event data using survival analysis.
  - Examples: time to death, time to tumor recurrence.

- Summaries of multiple event data.
  - Examples: the number of sex partners, number of seizures. Analysis techniques include Poisson and negative binomial regressions.

Graphical examples of outcome types in longitudinal studies



Possible hierarchies of units

- Time when the measurement if made (time-structured - longitudinal studies)
- Place (position, region) where the measurement is made (spatial data)
- Subunit on which the measurement is made (e.g., students within classrooms)
- Combinations

More complicated examples

- Measurements over time on various subunits (eyes within subjects)
- Measurements on subunits of subunits (hierarchical data structures), e.g., eyes within subjects within siblings
- Often the repeated measurements over time and/or place and/or subunit may also occur under different conditions (different treatments, covariate values, etc.) - cross-over trials.

Course Based mainly on Parametric Regression

- Emphasize estimating the association explanatory variables with an outcome variable.

- Could potentially examine any aspect of the distribution of the outcome, *Y*, conditional on the covariates, *X*.

- Most of the course concentrates on estimating how *X* affects the mean of *Y* (although other models are considered):

$$h\left(E[Y \mid X = x]\right) = \beta_0 + \beta_1 x$$

Types of Regression Models

- Continuous Outcome: Typically linear (*h(μ)=μ*):

$$E[Y \mid X = x] = \beta_0 + \beta_1 x$$

- Binary – Typically logistic, *h(μ)=log[μ/(1- μ)]*:

$$E[Y \mid X = x] = P(Y = 1 \mid X = x) = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 x)\}}$$

Types of Regression Models

- Counts – typically log-linear, $h(\mu)=log[\mu]$:

$$E[Y \mid X = x] = e^{\beta_0 + \beta_1 x}$$

- Often goal is to estimate β_0, β_1 (coefficients), which represent the parameters of interest.

Hazard Regression Models (Disease Incidence Data)

- In survival analysis of disease incidence data, the typical regression approach models the hazard

- The hazard $\lambda(t)$ is:

$$\lambda(t) = P(\text{fail "close" to time t} \mid \text{still at risk at time t})$$

- Possible model is:

$$\lambda(t \mid X = x) = \lambda_0(t)e^{\beta x}$$

Example 1
Multiple Event Data

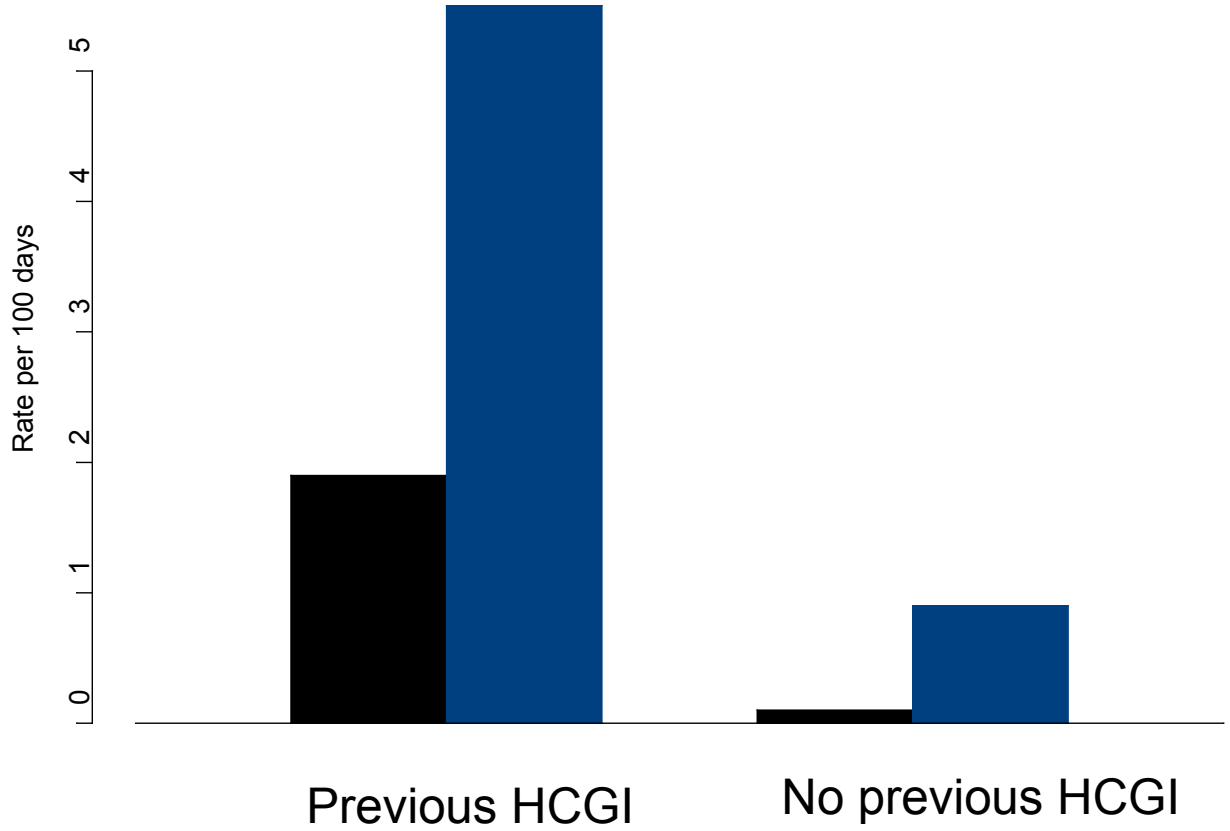
A randomized, controlled trial of an in-home drinking water intervention among HIV+ persons

- Pilot study of 50 HIV+ subjects who were randomized either active water filter or placebo device.
- Followed longitudinally and the number of highly credible gastro-intestinal (HCGI) events were recorded in (on average) a 6 month period.
- Purpose is to estimate the amount of HCGI attributable to drinking water among this population.

Table 1.2: Extract of Data from Study of dence of Gastrointestinal Symptoms

Id. Number	Date	hcgi	group
A7283	14780	-	6
A7283	14781	0	6
A7283	14782	0	6
A7283	14783	0	6
A7283	14784	0	6
A7283	14785	0	6
A7283	14786	0	6
A7283	14796	0	6
C1632	14738	-	7
C1632	14739	-	7
C1632	14740	-	7
C1632	14741	0	7
C1632	14742	0	7
C1632	14743	0	7
C1632	14744	1	7
C1632	14745	0	7
C1632	14746	0	7

Results of randomized water intervention among HIV+ persons



World Cup Soccer Data – Count Data

- The World Cup in soccer has been held every four years since 1930, except for 1942 and 1946 during and immediately following World War II.
- Data: the number of goals scored by a single team in every World Cup game played in 17 competitions.
- Outcome is a count.

World Cup Soccer Data

Table 1.5: Extract of Data from World Cup Soccer Results

Year	Continent	Goals	Teams
2002	0	0	4
2002	0	1	7
2002	0	2	5
2002	0	3	2
2002	0	4	1
2002	0	5	1
2002	1	0	16
2002	1	1	25
2002	1	2	10
2002	1	3	9
2002	1	4	1
2002	1	8	1
2002	2	0	17
2002	2	1	15
2002	2	2	12
2002	2	3	2

Example 2
Repeated Measures Data
Continuous Outcome

Longitudinal Data on HIV+ patients

- Deeks, et al. (1999) report the results from a longitudinal study of HIV-infected adults undergoing Highly Active Anti-Retroviral Therapy (HAART) at San Francisco General Hospital (SFGH).

- Patients were included in this analysis if they received at least 16 weeks of continuous therapy with an anti-retroviral regimen

- The following data was obtained during the initial review: date of birth, sex and length of previous exposure to each individual anti-retroviral agent.

- Once patients were identified, their medical records were reviewed every 3-4 months until November 1998.

- Plasma HIV RNA assays were performed using a branched DNA (bDNA) assay.

- Repeated and irregular measurements of CD4 and viral load (time-structured repeated measures)

- Data not always matched in time.

- Goal is to find how CD4 varies with viral load and how this pattern varies in the population

Sample of HIV+ Data

Table 1.1: Extract of Data from SFGH/HAART Stu

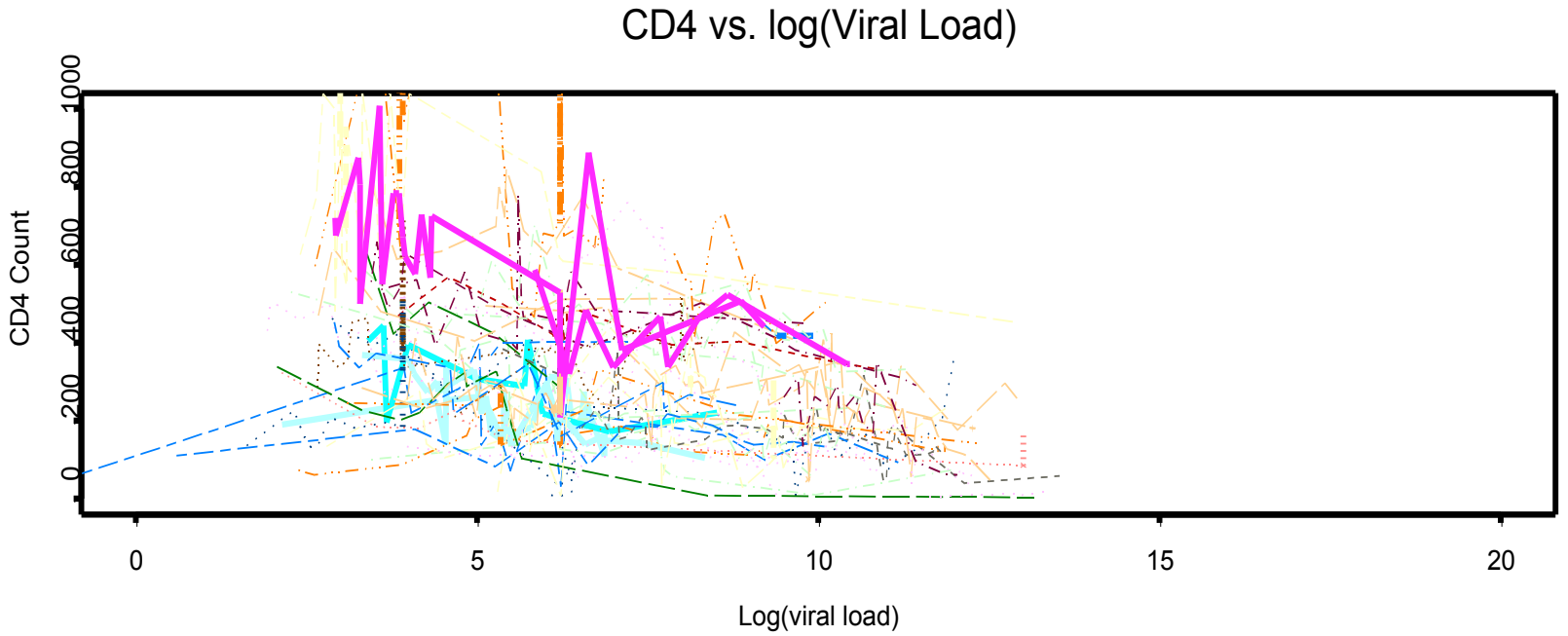
[1] Id. Number	days	CD4 count	log(viral load)	gender	age
1	39	45	2.70	1	32.0
1	137	119	5.22	1	32.0
1	147	113	.	1	32.0
1	179	74	5.20	1	32.0
1	187	95	.	1	32.0
1	298	137	3.87	1	32.0
1	335	.	5.07	1	32.0
1	354	167	5.14	1	32.0
1	411	.	4.66	1	32.0
1	1684	427	.	1	32.0
2	0	196	5.68	1	44.0
2	7	369	3.93	1	44.0
2	13	353	4.11	1	44.0
2	27	474	3.55	1	44.0
2	55	425	3.10	1	44.0
2	111	493	2.70	1	44.0
2	139	464	2.70	1	44.0
2	167	448	2.70	1	44.0
2	195	427	2.70	1	44.0

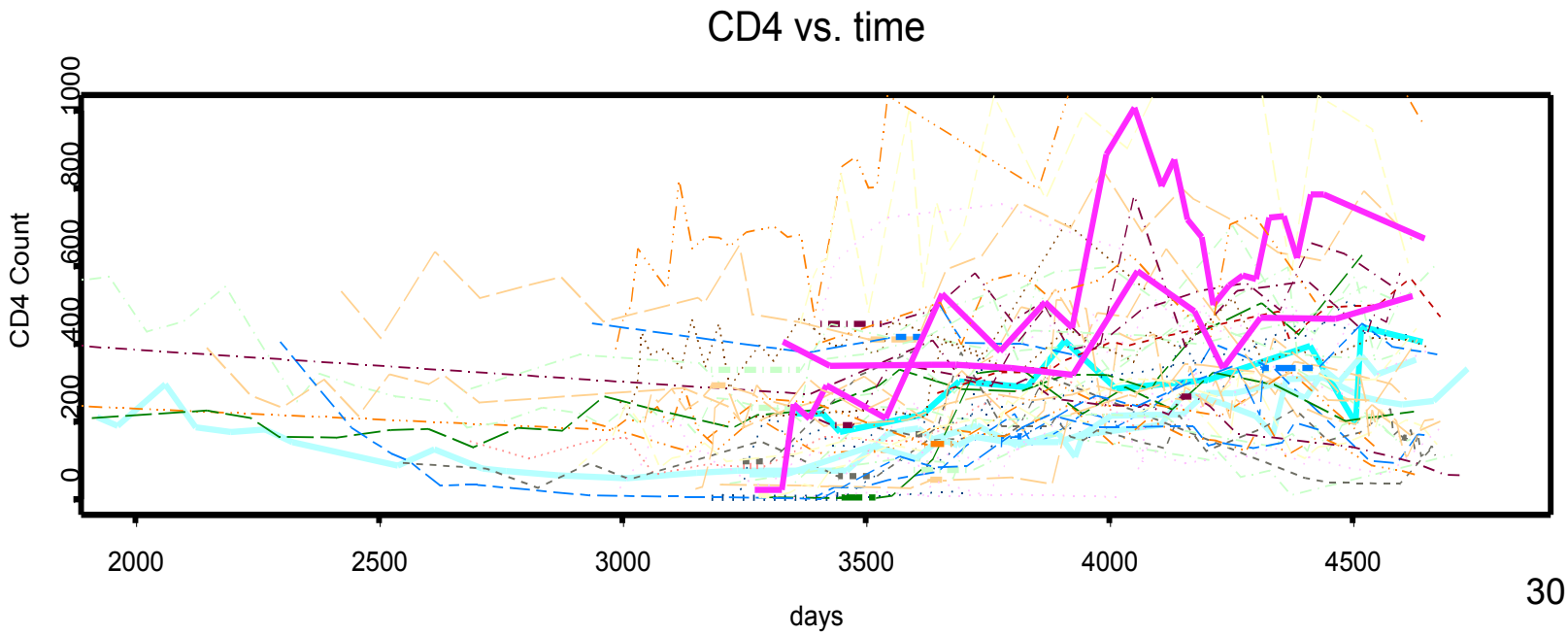
CD4 Count vs. Viral Load, cont.

- Possible Model:

$$Y_{ij}(t) = \beta_0 + \beta_{0i} + (\beta_1 + \beta_{1i}) X_{ij}(t - \delta) + e_{ij}$$

*i*th subject, *j*th measurement (at time *t*) time, $X_{ij(t-d)}$ is viral load at time *t-d*, $Y_{ij}(t)$ is the CD4 count at time *t*.


CD4 vs. log(Viral Load)
CD4 Count
0
200
400
600
800
1000
0
5
10
15
20
Log(viral load)


CD4 vs. time
CD4 Count
0
200
400
600
800
1000
2000
2500
3000
3500
4000
4500
days

Example 3
Repeated Measures Data
Binary Outcome

The Effect of Drug and Alcohol Use on Teenage Sexual Activity

- Minnis & Padian (2001) conducted a longitudinal study of teenagers in San Rafael, California to investigate the association between drug and alcohol use and sexual activity on the same day.

- Participants were asked to keep track of their activities over approximately one month and binary indicator variables were created to show whether drug/alcohol use and/or sexual activity were reported for each 24 hour period.

The Effect of Drug and Alcohol Use on Teenage Sexual Activity

- Data is available for 109 teenagers for whom information on 1 to 33 different days are available.
- The average number of longitudinal observations is 16, with the total number of data points (that is, teenager-days) equal to 1,708.

Extract of Teenage Drugs and Sex Data

Table 1.3: Extract of Data from Teenage Survey on Dr Sexual Activity

Id. Number	Date	Drug/Alcohol Use	Sexual Activity
10122	03 Jun 98	yes	no
10123	04 Jun 98	no	no
10123	05 Jun 98	no	no
10123	06 Jun 98	yes	no
10123	07 Jun 98	no	no
10123	08 Jun 98	no	no
10123	09 Jun 98	no	no
10123	12 Jun 98	no	no
10123	14 Jun 98	yes	no
10123	16 Jun 98	no	no
10123	17 Jun 98	no	no
10123	18 Jun 98	no	yes
10123	19 Jun 98	no	no
10123	20 Jun 98	no	no
10123	21 Jun 98	no	no
10123	23 Jun 98	no	no
10123	25 Jun 98	no	yes
10123	28 Jun 98	no	no
10123	29 Jun 98	no	yes

Example 4
Time to Event Data

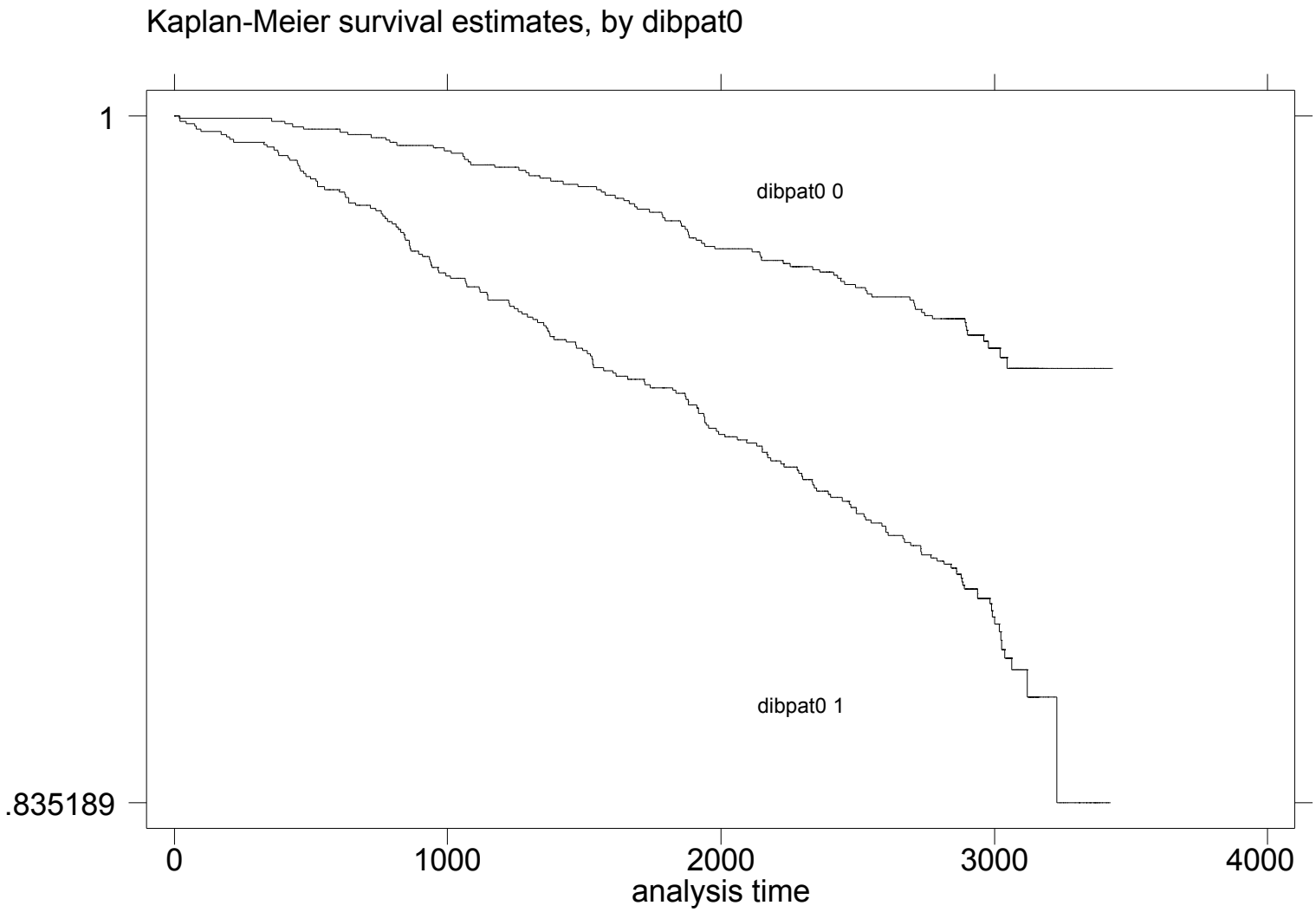
Western Collaborative Group Study

- Collected follow-up data on 3,154 employed men from 10 Californian companies (1960-61).
- Aged 39-59 years old at baseline
- Looked for onset of CHD for about 9 years
- Risk factors measured: smoking, blood pressure, cholesterol,weight, behavior type
- 257 CHD “events”

Western Collaborative Group Study

id	age0	height0	weight0	chol0	behpat0	ncigs0	chd69	time169
2001	49	73	150	225	2	25	0	1664
2002	42	70	160	177	2	20	0	3071
2003	42	69	160	181	3	0	0	3071
2004	41	68	152	132	4	20	0	3064
2005	59	70	150	255	3	20	1	1885
2006	44	72	204	182	4	0	0	3102
2007	44	72	164	155	4	0	0	3074
2008	40	71	150	140	2	0	0	3071
2009	43	72	190	149	3	25	0	3064
2010	42	70	175	325	2	0	0	1032
2011	53	69	167	223	2	25	0	3091
2013	41	67	156	271	2	20	0	3081
2014	50	72	173	238	1	50	1	1528
2017	43	72	180	189	3	30	0	3072

Western Collaborative Group Study



Example 5
Ecological Time Series

Leptorospirosis and Climate

- Leptospirosis is a **bacterial disease** that affects humans and animals.

- In humans it causes a wide range of symptoms with around 5–10% of infected individuals suffering severe forms of the disease, and, on rare occasions, death.

- **Outbreaks of leptospirosis** are usually caused by exposure to water contaminated with the urine of infected animals, typically following heavy rainfall with subsequent sewer flooding.

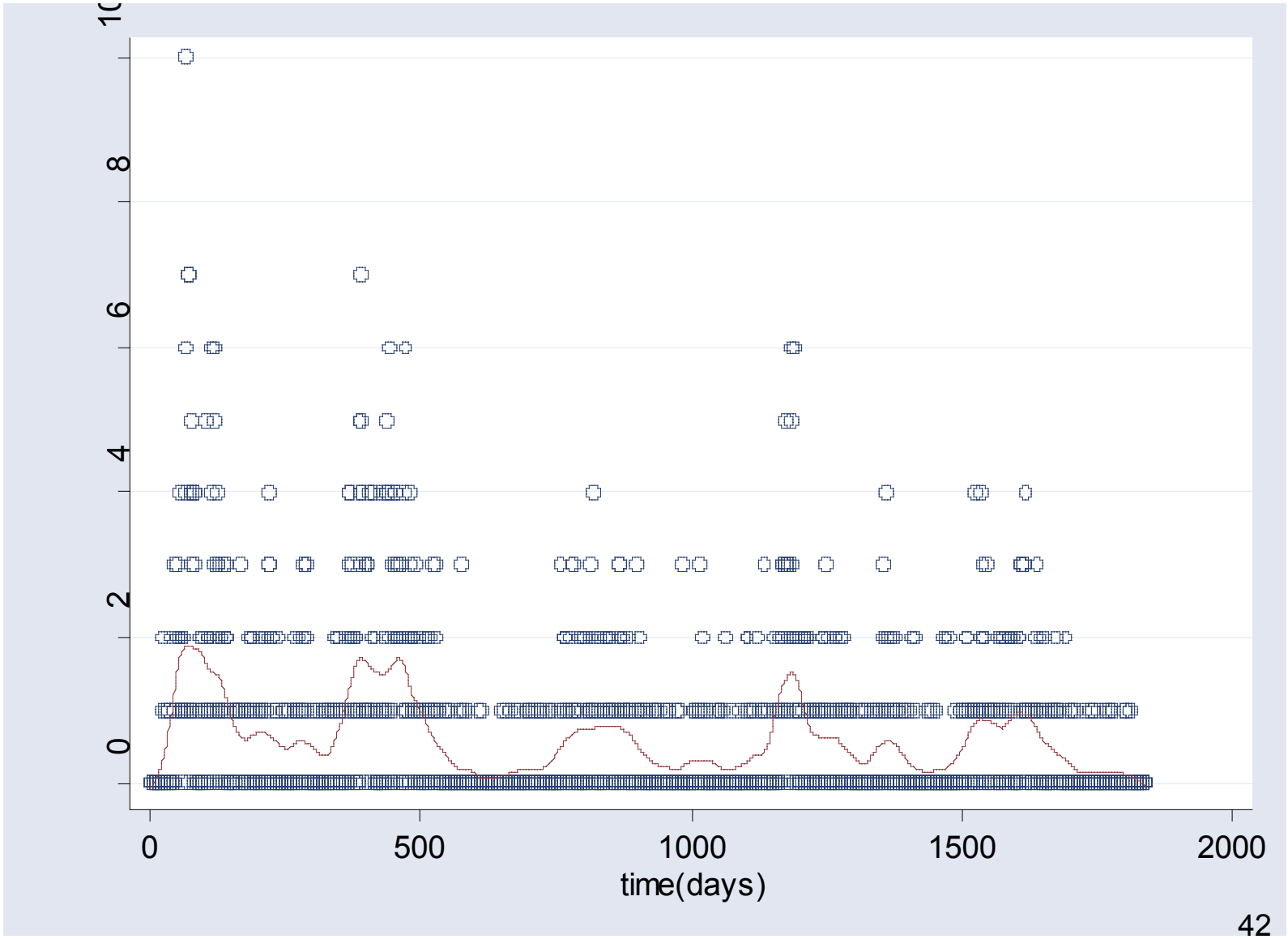
- Urban outbreaks in large Latin American city slums are assumed to result from **poor sanitation** infrastructure and proliferation of **rodent populations**.

Leptorospirosis and Climate

- The data used here arose from **surveillance data** in an infectious disease hospital in Salvador, Brazil, an institution that accounts for 95% of case notifications in the city (Flannery et al., 2001).

- In addition, **meteorological information** on daily rainfall, temperature (maximum, minimum, and average), and relative humidity for the same period were also collected.

- One goal for data analysis is **estimation of the lag time between high rainfall days and days of high case counts**, providing insight into the disease's incubation period in addition to suggesting appropriate time periods for possible intervention after periods of heavy rain.


10
8
6
4
2
0
0
500
1000
1500
2000
time(days)