Analysis of Failure Time Data from Screening Studies with Missing Observations

Dianne M. Finkelstein
MGH Biostatistics Center
Harvard School of Public Health

August 2003

Missing Data in Screening Studies

- Patients are monitored for occurrence of events.
- Event can only be detected by a clinical exam or lab test at clinic visit.
  - HIV test for presence of antibody
- Some patients miss exams and we only know the event took place between the subject's last negative and first positive screening time.
- Data set consists of overlapping intervals in which failures occurred.

Complex Missing Screening Data

- Often there are two events of interest in the study
- First event marks onset of disease–antibody present –interval censored
- Second event marks a progression of the disease –symptomatic stage
  - Could be Exact/right censored– such as diagnosis of Opportunistic infection indicating AIDS
  - Could be interval censored as well–CD4<200 indicating AIDS
- Sometimes interest is focused on the time between the two events–latency–where at the first (and possibly second) are interval censored
- In addition time-varying covariate information required for regression analysis may be missing

Example: CMV shedding in the blood and urine of HIV patients

- Patients were participating in a PCP prophylaxis trial
- Observational substudy monitored patients every 6 months for CMV shedding in the blood and urine
- All shedding events were censored into the interval between last negative and first positive screen.
- Shedding is asymptomatic but often precedes CMV Retinitis which results in blindness
- Time of CMV Retinitis diagnosis was recorded (exact/right censored)
- Analyses of interest: What is the time from shedding to disease?
- Data given in Betensky and Finkelstein **Statistics in Medicine** 1999

Screening Diabetics for Proteinuria (Kidney Disease)

- Patients were monitored every 6 months for excess urinary albumin excretion indicating nephropathy
- Microalbuminuria (trace of albumin) indicates early kidney disease
- Proteinuria (higher levels of albumin) marks progressive loss of renal function
- Question of interest
  - Glycohemoglobin A1C is an indicator of poor diabetes control
  - Does A1C predict progression to Proteinuria?
- Problems with the data
  - When screening visit is missed, time of Microalbuminuria and Proteinuria are interval censored.
  - When a visit is missed, A1C (the covariate) is also missing

Analysis of CMV Latency

- Patients monitored every 6 months for CMV
  - CMV shedding in blood and urine
  - Onset of CMV Retinitis
- First indication of shedding in urine
  - 40 left censored
  - 70 interval censored
  - 67 right censored
- First indication of shedding in blood
  - 5 left censored
  - 22 interval censored
  - 150 right censored
- CMV Diagnosis
  - 33 exact
    5 had last observation CMV Negative
  - 144 right censored
- What is the time from first shedding until CMV diagnosis?

Estimation of Latency: DeGruttola and Lagakos 1989 (Biometrics)

- Nonparametric estimate from self-consistency equations
- Noted that cannot transform data and apply univariate methods
- Contribution to likelihood from subject i

$$\sum_j \sum_k \alpha_{jk}^i w_j f_k \tag{1}$$

- where $\alpha_{jk}^i = 1$ if observed data for i^{th} subject consistent with infection at j and latency of k
- w_j is density for infection
- f_k is density for latency
- Assumes latency independent of infection time.
- If both j and k indexed f, would allow dependence
- If impute infection time, likelihood separates.
- Issue: Independence may not always be valid

Estimation of Latency Distribution Assuming Dependence on Infection Time

- Infection T_1, Disease onset T_2, Latency $T_2 - T_1$
- Two approaches:
  1. Can factor joint distribution of infection and latency allowing dependence:

$$Pr(T_1, T_2 - T_1) = Pr(T_1) \cdot Pr(T_2 - T_1 | T_1) \quad (2)$$

  2. Can directly estimate the joint distribution of infection and disease onset, $Pr(T_1, T_2)$ and calculate latency from the convolution.

Estimation of Bivariate Failure

- Betensky and Finkelstein (SIM 1999)
- Showed that support for MLE is contained in a set of rectangles of the plane
- Could now be considered as a univariate problem by indexing the rectangles of support, j
- Likelihood for infection T_1 and disease onset T_2

$$\prod_i \sum_j \alpha_j^i g_j \tag{3}$$

- g_j is probability associated with j^{th} square
- $\alpha_j^i = 1$ if observed infection and disease for i^{th} subject could be in j^{th} square
- Becomes a generalization of Turnbull 1976.

Bivariate Estimate

T_1 Left	T_1 Right	T_2 Left	T_2 Right	Probability
1	7	10	10	0.122770066
9	9	11	11	0.076130347
9	9	12	12	0.050269077
1	7	9	9	0.023584944
1	7	14	14	0.046365059
8	8	9	9	0.002747089
9	9	14	14	0.061788326
8	8	14	14	0.001785231
10	10	11	11	0.065641243
10	10	12	12	0.113798321
11	11	14	14	0.060373420
12	12	13	13	0.120690554
12	12	12	12	0.021643962
11	11	15	15	0.016371399
13	13	14	14	0.102425363
13	13	15	15	0.053398388
14	14	15	15	0.032055336
8	8	8	8	0.013954193
15	15	15	16	0.014207681

Calculation of Latency

- Latency is the convolution of T_2 and T_1
- If data were complete, $\hat{g}_j$ could be calculated as proportion of j^{th} square within the grid
- With interval censored data, support for distribution g_j is on disjoint rectangles.
- The distribution is indeterminate for squares within these rectangles–non-identifiability
- Cannot directly calculate $T_2 - T_1$ for each observation, as T_1 can be an interval and T_2 could be right censored
  - Example: Probability for $(1, 7] \quad (10]$ is .1228
  - Latency could be 8 if infection in $(1, 2] \quad \ldots$
  - Latency could be 1 if infection in $(6, 7]$
- Solution: Assume uniform distribution of the g_j over these squares (Note that this is still an MLE)
- Calculation of an MLE for latency is now simple
  - Probability of latency 8 is $\frac{1}{8} \cdot .1228 = .0154 \quad \ldots$
  - Probability of latency 1 is $\frac{1}{8} \cdot .1228$

Generalizing Degruttola et al (1989) to Allow Dependence

- Degruttola et al (1989) factored joint distribution
  $Pr(T_1, T_2 - T_1) = Pr(T_1) \cdot Pr(T_2 - T_1)$

$$L = \prod_i \sum_j \sum_k \alpha_{jk}^i w_j f_k \tag{4}$$

- Instead factor the joint distribution:
  $Pr(T_1, T_2 - T_1) = Pr(T_1) \cdot Pr(T_2 - T_1 | T_1)$

$$L = \prod_i \sum_j \sum_k \alpha_{jk}^i w_j f_{jk} \tag{5}$$

  - where α_{jk} is indicator could be $T_1 = j, T_2 - T_1 = k$
  - w_j is infection at j
  - f_{jk} is latency distribution at k given infection at j

- Frydman (1995) proposed

Extend To Incorporate Covariates on Infection and Latency

- Likelihood factored as before, but include covariate z

$$L = \prod_i \sum_j \sum_k \alpha_{jk}^i w_j(z) f_{jk}(z) \tag{6}$$

- Model dependence of infection, T_1 on covariates Z

$$\text{Logit} W_j(z) = \mu_j + \beta_1 z \tag{7}$$

  where W_j is CDF for infection

- Model dependence of latency $T_2 - T_1$ on covariates and infection time

$$\text{Logit} F_{jk}(z) = \mu_k + \beta_2 z + \beta_3 \gamma(j) \tag{8}$$

  where F_{jk} is CDF for latency

- Work in progress

Principles of the Regression Methodology

- Discretize time by categorizing simultaneously on both dimensions (infection and latency)
- May have to group data
- Multinomial model
- Use E-M algorithm because it simplifies when data are complete
- Non-parametric methods require large number of parameters
- Computationally intensive
- Simplified if points of positive mass are known
- Issues of identifiability–mild parametric assumptions

Missing Failure Time Observations and Time-Varying Covariates

- Regression with missing outcomes and covariates
- Analysis of the relationship of recent A1C on risk for proteinuria in patients with microalbuminuria
- Joint distribution for A1C Z and Proteinuria T modeled as

$$L(t, z) = g(t|z) \cdot m(z) \tag{9}$$

- $m(z)$ longitudinal model for A1C–random effects model
- $g(t|z)$ logistic model for Proteinuria as a function of previous A1C
- Use Pooling Repeated Observations (PRO) method to model person-exam risk
  - Cupples et al (1988) SIM
  - Asymptotically equivalent to grouped Cox model

Multiple Imputation

- Used an adaptation of the Predictive Mean Matching method (Heitjan and Little JRSS C 1991)
- Missing progression imputed as follows:
  - Fit logistic model on complete data to get $\hat{\beta}$
  - For all subjects (including missing and complete), use $\hat{\beta}$ to get predicted probability of progression T for each subject
  - Divide the sample into deciles by these probabilities
  - For each bin, sample with replacement from complete observations to create a bootstrap sample equal to number of complete in that bin.
  - For each subject with missing outcome in the bin, sample with replacement from bootstrap sample
  - Combine these imputed and complete data to get one imputation.
- Produce 5 imputed samples
- Calculate new estimates for β as in Rubin (1986)

Handling Missing Time-Varying Covariate (A1C)

- Fit random effects model $m(z)$ for A1C
- Create decile bins based on predicted A1C
- Create bootstrap sample of complete subjects in each bin
- Select imputed data for missing Z as before for T
- Produce one imputed set of Z
- Include these in the logistic model to predict progression as described above.

Analysis of Proteinuria in Diabetes

- 366 Subjects with maximum of 4 biannual visits
- Initial analysis selected on the complete data:
  - Only had 929 subject-visits
  - Odds Ratio 8.103 (3.1, 21.1)
- Next investigator asked to have A1C "filled in"
  - A1C tracks, so suggested that we use the last non-missing observation to complete the missing data on A1C (LOCF method).
  - Only use person-exams with complete failures
  - 975 person-visits
  - Odds Ratio: 6.0 (2.6, 13.8)
- We applied Using Multiple Imputation:
  - 1464 person-visits
  - Odds Ratio 4.3 (1.6, 11.3)

Discussion

- Missingness or Censoring could be dependent
  - Finkelstein, Schoenfeld and Goggins 2002 handled dependent interval censoring.
  - Need to generalize to latency
  - Generalize Multiple Imputation method
- Computational/asymptotic issues
  - Algorithm is slow–convergence issues
  - Number of parameters is large and increasing
- Multiple Imputation, GEE, etc have moved into the realm of non-statisticians.

Acknowledgments to Collaborators

- Rebecca Betensky
  Harvard SPH
- David Schoenfeld
  Mass General Hospital
- Linda Ficociello
  Joslin Diabetes Center

References

Interval Censored Data

- Turnbull, B.W., The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B*, **38**, 290-295, 1976.
- Lindsey, JC. and Ryan, Louise M., Methods for Interval-censored Data *Statistics in Medicine*, 1998, **17**: 219–238
- Finkelstein, DM. Goggins, WB, Schoenfeld, DA, Analysis of failure time data with dependent interval censoring. *Biometrics*, 2002; 58(2). 298-304.

Doubly Censored (Latency) Data

- De Gruttola V. and Lagakos, S.W., 'Analysis of doubly-censored survival data, with application to AIDS', *Biometrics*, **45**, 1-11 (1989).
- Kim, M.Y., De Gruttola, V., and Lagakos, S.W., 'Analyzing doubly censored data with covariates, with applications to AIDS', *Biometrics*, **49**, 13-22 (1993).
- Goggins W, Finkelstein DM and Zaslavsky A. Applying the Cox Proportional Hazards Model when the Change Time of a Binary Time-Varying Covariate is Interval-Censored. Biometrics 1999;55: 445-451.
- Betensky R, and Finkelstein DM. A non-parametric maximum likelihood estimator for bivariate interval censored data, *Statistics in Medicine* 1999 Nov 30;18(22):3089-100.
- Frydman, H, Semiparametric estiation in a three state duration dependent Markov Model from interval-censored observations with applications to AIDS data, *Biometrics* 1995, **51** 502-511.

Multiple Imputation

- Heitjan D and Little R. Multiple Imputation for the Fatal Accident Reporting System. *Applied Statistics* 1991: 40(1), 13-29.
- Rubin D. Multiple Imputation for Nonresponse in Surveys. Wiley, Johnson, and Sons, Inc. New York, 1987.
- Rubin D. Multiple Imputation after 18+ Years. *Journal of the American Statistical Association* 1996: 91(434), 473-489.

Pooled Logistic Regression

- Cupples L.A., D'Agostino, R.B., Anderson, K. and Kannel, W.B., Comparison of baseline and repeated measure covariate techniques in the Framingham Heart Study, *Statistics in Medicine* 1988, **7**, 205-218.

- D'Agostino, R., Lee, ML, Belanger, A, Cupples, A, Anderson, K, Kannel, W, Relation of pooled logistic regression to time dependent Cox regresion analysis: the Framingham Heart Study, *Statistics in Medicine* 1990, **9**, 1501-15.

http://hedwig.mgh.harvard.edu/biostatistics/index.html