Longitudinal modeling when the response and time-dependent covariate(s) are measured at distinct time points

Joel A. Dubin (University of Waterloo)
and Xiaoqin Xiong (UW),
with acknowledgment to
Dr. George Kaysen and Dr. Patrick Romano (UCD)

CANNeCTIN Biostatistics Methodology
Videoconference Seminar Series

Jan 08, 2010

Outline

Outline

- Some history — an earlier biomedical/biostatistical problem.

Outline

- Some history — an earlier biomedical/biostatistical problem.
- The current problem — longitudinal association between a continuous and binary response measured at different time points.

Outline

- Some history — an earlier biomedical/biostatistical problem.
- The current problem — longitudinal association between a continuous and binary response measured at different time points.
- Method and analysis results.

Outline

- Some history — an earlier biomedical/biostatistical problem.
- The current problem — longitudinal association between a continuous and binary response measured at different time points.
- Method and analysis results.
- Simulation results.

Outline

- Some history — an earlier biomedical/biostatistical problem.
- The current problem — longitudinal association between a continuous and binary response measured at different time points.
- Method and analysis results.
- Simulation results.
- Discussion.

Initial problem — nephrology study

Initial problem — nephrology study

- Observational study of hemodialysis patients (n=35).

Initial problem — nephrology study

- Observational study of hemodialysis patients (n=35).
- Measurements taken longitudinally for five proteins:

Initial problem — nephrology study

- Observational study of hemodialysis patients (n=35).
- Measurements taken longitudinally for five proteins:
  - negative APPs: alb, trf

Initial problem — nephrology study

- Observational study of hemodialysis patients (n=35).
- Measurements taken longitudinally for five proteins:
  - negative APPs: alb, trf
  - positive APPs: crp, cer, aag

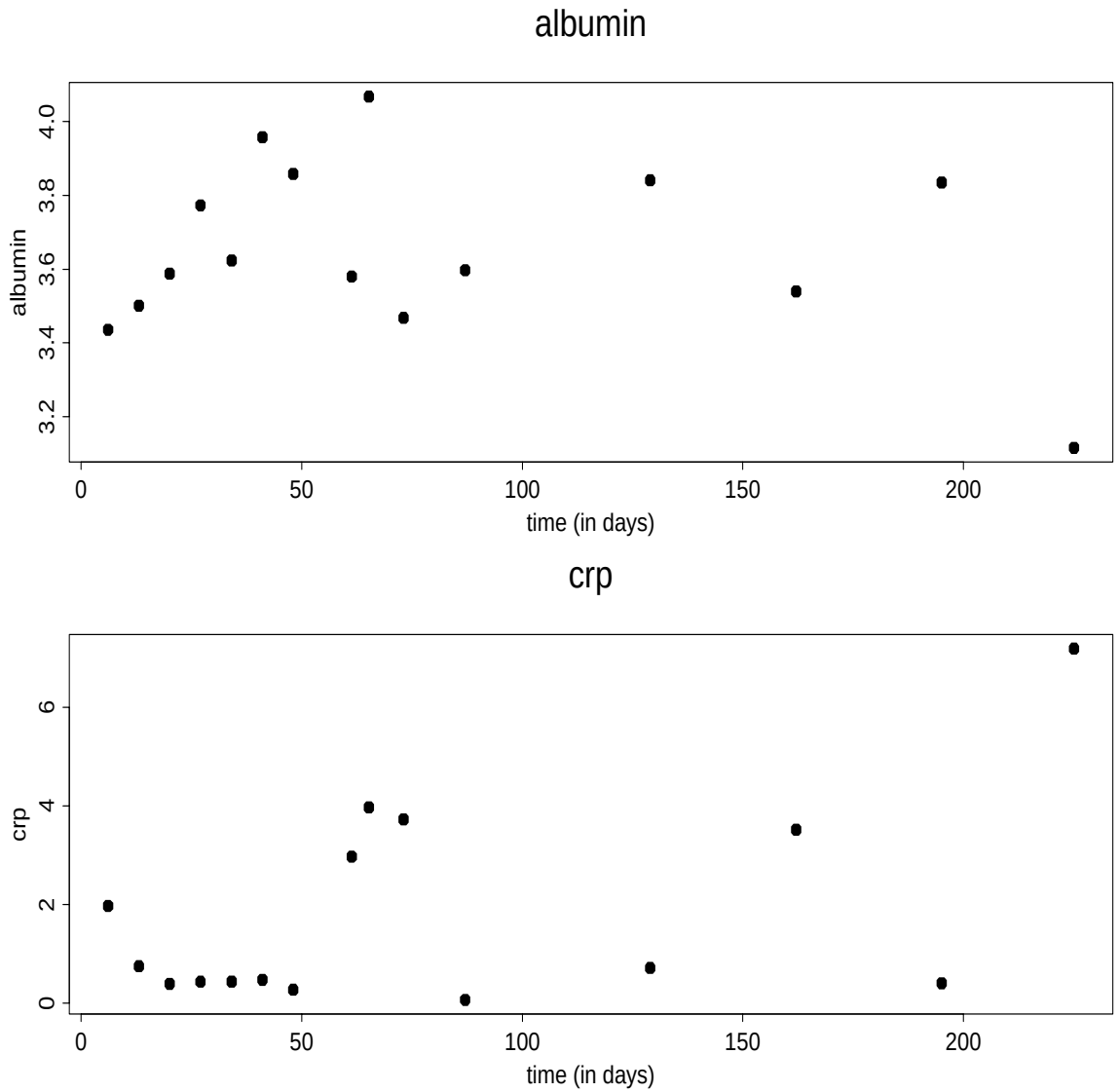
Initial problem — nephrology study

- Observational study of hemodialysis patients (n=35).
- Measurements taken longitudinally for five proteins:
  - negative APPs: alb, trf
  - positive APPs: crp, cer, aag
- Design is unbalanced, with between 12 and 18 multivariate measurements per patient.

Initial problem — nephrology study

- Observational study of hemodialysis patients (n=35).
- Measurements taken longitudinally for five proteins:
  - negative APPs: alb, trf
  - positive APPs: crp, cer, aag
- Design is unbalanced, with between 12 and 18 multivariate measurements per patient.
- Goal of initial analysis was to determine how proteins are correlated over time, including consideration of class and lagged effects.

Fig. 1: Observed values for albumin and crp for one random patient



Summary of approach to older problem

Summary of approach to older problem

- called **dynamical correlation**, we developed a curve-based approach to measure association between pairwise longitudinal proteins

Summary of approach to older problem

- called **dynamical correlation**, we developed a curve-based approach to measure association between pairwise longitudinal proteins
- extensions were provided to look at lagged relationships and derivatives; also, multivariate techniques were applied to look at correlation between classes of proteins

Summary of approach to older problem

- called **dynamical correlation**, we developed a curve-based approach to measure association between pairwise longitudinal proteins
- extensions were provided to look at lagged relationships and derivatives; also, multivariate techniques were applied to look at correlation between classes of proteins
- more details in Dubin and Müller (2005)

Summary of approach to older problem

- called **dynamical correlation**, we developed a curve-based approach to measure association between pairwise longitudinal proteins
- extensions were provided to look at lagged relationships and derivatives; also, multivariate techniques were applied to look at correlation between classes of proteins
- more details in Dubin and Müller (2005)
- one limitation: a particular high correlation between two proteins for a given individual said nothing about that person's health status

A case of "augmented data"

A case of "augmented data"

- Interest arose to investigate the relationship between the behavior of the proteins and certain events of interest for the hemodialysis patients (such as infection and access events). So, new data was obtained from a large subset of patients from the earlier study.

A case of "augmented data"

- Interest arose to investigate the relationship between the behavior of the proteins and certain events of interest for the hemodialysis patients (such as infection and access events). So, new data was obtained from a large subset of patients from the earlier study.
- Actually, this "new data" was existing data for the patients, from their chart records, which were not collected as part of the initial study.

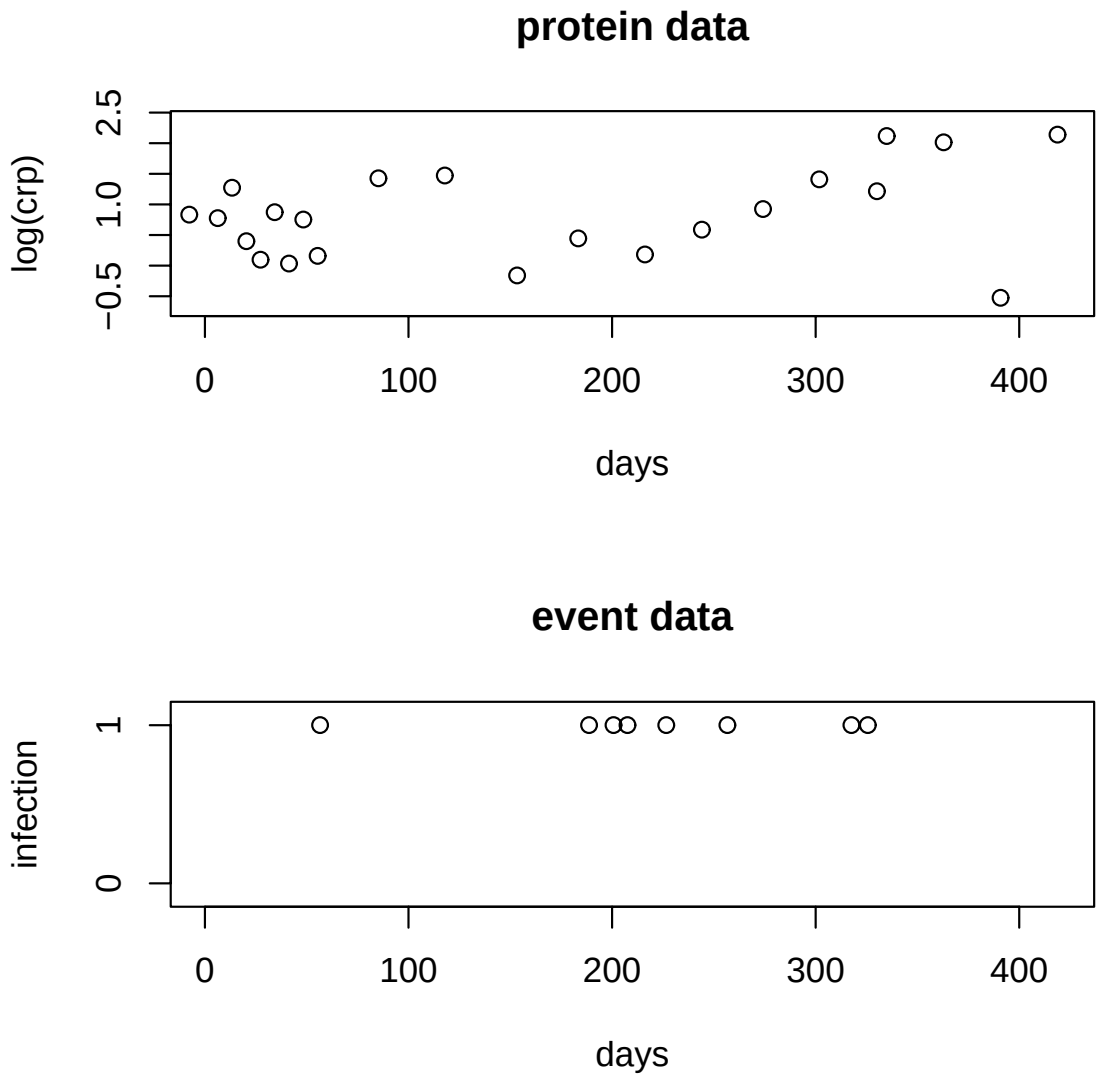
A case of "augmented data"

- Interest arose to investigate the relationship between the behavior of the proteins and certain events of interest for the hemodialysis patients (such as infection and access events). So, new data was obtained from a large subset of patients from the earlier study.
- Actually, this "new data" was existing data for the patients, from their chart records, which were not collected as part of the initial study.
- Key questions: Is a rise/decline in one of the proteins associated with a contemporaneous event, and can we detect if one typically precedes the other?

A case of "augmented data"

- Interest arose to investigate the relationship between the behavior of the proteins and certain events of interest for the hemodialysis patients (such as infection and access events). So, new data was obtained from a large subset of patients from the earlier study.
- Actually, this "new data" was existing data for the patients, from their chart records, which were not collected as part of the initial study.
- Key questions: Is a rise/decline in one of the proteins associated with a contemporaneous event, and can we detect if one typically precedes the other?
- Key problem: the days of the chart data did not coincide with the days of the protein data.

Fig. 2: Observed values for protein and event for one patient



Set-up for modeling

Set-up for modeling

- Let $Y_{i,j}$ be binary health event observed for patient i at time j, where $j = 1, 2, ..., n_i^{(Y)}$.

Set-up for modeling

- Let $Y_{i,j}$ be binary health event observed for patient i at time j, where $j = 1, 2, ..., n_i^{(Y)}$.
- Let $X_{i,k}$ be continuous protein measurement for patient i observed at time k, for $k = 1, 2, ..., n_i^{(X)}$, where, typically, the times represented by $k \neq$ the times represented by j and $n_i^{(X)} \neq n_i^{(Y)}$.

Set-up for modeling

- Let $Y_{i,j}$ be binary health event observed for patient i at time j, where $j = 1, 2, ..., n_i^{(Y)}$.
- Let $X_{i,k}$ be continuous protein measurement for patient i observed at time k, for $k = 1, 2, ..., n_i^{(X)}$, where, typically, the times represented by $k \neq$ the times represented by j and $n_i^{(X)} \neq n_i^{(Y)}$.
- Need some type of smoothing to allow for longitudinal modeling of Y on X for $N = 53$ patients.

Set-up for modeling

- Let $Y_{i,j}$ be binary health event observed for patient i at time j, where $j = 1, 2, ..., n_i^{(Y)}$.
- Let $X_{i,k}$ be continuous protein measurement for patient i observed at time k, for $k = 1, 2, ..., n_i^{(X)}$, where, typically, the times represented by $k \neq$ the times represented by j and $n_i^{(X)} \neq n_i^{(Y)}$.
- Need some type of smoothing to allow for longitudinal modeling of Y on X for $N = 53$ patients.
- A simple idea: bin (X, Y) in equidistant units of time; then take unweighted or weighted average (or sum) of variables inside each bin.

Set-up for modeling

- Let $Y_{i,j}$ be binary health event observed for patient i at time j, where $j = 1, 2, ..., n_i^{(Y)}$.
- Let $X_{i,k}$ be continuous protein measurement for patient i observed at time k, for $k = 1, 2, ..., n_i^{(X)}$, where, typically, the times represented by $k \neq$ the times represented by j and $n_i^{(X)} \neq n_i^{(Y)}$.
- Need some type of smoothing to allow for longitudinal modeling of Y on X for $N = 53$ patients.
- A simple idea: bin (X, Y) in equidistant units of time; then take unweighted or weighted average (or sum) of variables inside each bin.
- Resulting data will be $(X_{i,m}, Y_{i,m})$, where $m = 1, 2, 3, ..., n_i^{(X,Y)}$.

Modeling approach

Modeling approach

- For each patient, we initially take the sum of events ($Y_{i,m}$) within each bin, and assume that conditional on $X_{i,m}$ and a sole subject-specific random effect b_i, these events are distributed as Poisson(μ_i).

Modeling approach

- For each patient, we initially take the sum of events ($Y_{i,m}$) within each bin, and assume that conditional on $X_{i,m}$ and a sole subject-specific random effect b_i, these events are distributed as Poisson(μ_i).

- We will enter the now-aligned longitudinal measurements for the event and protein into a generalized linear mixed effects model for a count response. We also consider zero-inflated extensions.

Modeling approach

- For each patient, we initially take the sum of events ($Y_{i,m}$) within each bin, and assume that conditional on $X_{i,m}$ and a sole subject-specific random effect b_i, these events are distributed as Poisson(μ_i).

- We will enter the now-aligned longitudinal measurements for the event and protein into a generalized linear mixed effects model for a count response. We also consider zero-inflated extensions.

- Specifically, we fit a Poisson model with a normal random effect, and a mixed ZIP model with random effects for possibly both parts of the mixture; for model fitting, we used the NLMIXED procedure in SAS, which uses AGQ for approximating the likelihood.

Modeling approach (cont.)

Modeling approach (cont.)

- As for binning, we took two approaches. The first utilized the entire time course of data for each subject (up to 1 1/2 years), where the protein values were taken roughly every seven days for the first seven weeks under observation and every month thereafter. Events could be measured whenever the patients took their dialysis treatment, which was three times per week. Binning choices here included 30 and 45-day bins.

Modeling approach (cont.)

- As for binning, we took two approaches. The first utilized the entire time course of data for each subject (up to 1 1/2 years), where the protein values were taken roughly every seven days for the first seven weeks under observation and every month thereafter. Events could be measured whenever the patients took their dialysis treatment, which was three times per week. Binning choices here included 30 and 45-day bins.
- The second binning approach focused only on the first seven weeks of follow-up data, as this allowed for closer correspondence in time between event recording and protein measurement. We considered 7-day and 10-day bins for this subset dataset.

Modeling approach (cont.)

- As for binning, we took two approaches. The first utilized the entire time course of data for each subject (up to 1 1/2 years), where the protein values were taken roughly every seven days for the first seven weeks under observation and every month thereafter. Events could be measured whenever the patients took their dialysis treatment, which was three times per week. Binning choices here included 30 and 45-day bins.
- The second binning approach focused only on the first seven weeks of follow-up data, as this allowed for closer correspondence in time between event recording and protein measurement. We considered 7-day and 10-day bins for this subset dataset.
- An important question to answer was not only "is there a contemporaneous association between event occurrence and protein levels?", but "is there a lagged association such that there is plausibility that one "process" precedes the other?".

Model results following binning

Model results following binning

Following results are from the mixed zero-inflated Poisson model with two binning approaches (30-day and 7-day)

protein: crp event: infection

Model results following binning

Following results are from the mixed zero-inflated Poisson model with two binning approaches (30-day and 7-day)

protein: crp event: infection

bin approach	lag	$\hat{\beta}$	SE	p-value	RE(1) s.d.	RE(2) s.d.
30-day	0	0.533	0.102	$<$ **.0001**	0.714	1.071
	-1	0.212	0.159	0.188	0.717	0.774
	1	0.165	0.108	0.134	0.390	1.188
7-day	0	0.490	0.317	0.128	0.972	
	-1	0.350	0.344	0.314	0.932	
	1	1.115	0.364	**0.004**	1.003	

Note: lag of -1 means crp is lagged behind infection occurrence, and lag of 1 means infection occurrence is lagged behind crp.

Note: RE(1) refers to random effect from log-linear piece, and RE(2) refers to random effect from logit (zero mixing) piece.

Model results following binning

Following results are from the mixed zero-inflated Poisson model with two binning approaches (30-day and 7-day)

protein: crp event: infection

bin approach	lag	$\hat{\beta}$	SE	p-value	RE(1) s.d.	RE(2) s.d.
30-day	0	0.533	0.102	$<$ **.0001**	0.714	1.071
	-1	0.212	0.159	0.188	0.717	0.774
	1	0.165	0.108	0.134	0.390	1.188
7-day	0	0.490	0.317	0.128	0.972	
	-1	0.350	0.344	0.314	0.932	
	1	1.115	0.364	**0.004**	1.003	

Note: lag of -1 means crp is lagged behind infection occurrence, and lag of 1 means infection occurrence is lagged behind crp.

Note: RE(1) refers to random effect from log-linear piece, and RE(2) refers to random effect from logit (zero mixing) piece.

Simulation setup

Simulation setup

- we based the setup on the hemodialysis data

Simulation setup

- we based the setup on the hemodialysis data
- we initiated the data generation by simulating normally distributed values for $X_{i,j}$, conditional on a random intercept, then generated $Y_{i,j}$ as Poisson, conditional on $X_{i,j}$ and its own random intercept.

Simulation setup

- we based the setup on the hemodialysis data
- we initiated the data generation by simulating normally distributed values for $X_{i,j}$, conditional on a random intercept, then generated $Y_{i,j}$ as Poisson, conditional on $X_{i,j}$ and its own random intercept.
- we then imposed a mismatch mechanism, such that all, some, or none of the $X_{i,j}$ and $Y_{i,j}$ were observed on the same days across all subjects

Simulation setup

- we based the setup on the hemodialysis data
- we initiated the data generation by simulating normally distributed values for $X_{i,j}$, conditional on a random intercept, then generated $Y_{i,j}$ as Poisson, conditional on $X_{i,j}$ and its own random intercept.
- we then imposed a mismatch mechanism, such that all, some, or none of the $X_{i,j}$ and $Y_{i,j}$ were observed on the same days across all subjects
- we considered factors such as mismatch rate, autocorrelation and within-subject variability when generating the $X_{i,j}$, levels of between-subject variability of $X_{i,j}$ and $Y_{i,j}$, bin size, and number of obs within a fixed bin size

Highlights of simulation results

Highlights of simulation results

- Not surprisingly, this method works well when we have matched data (matched at the same time points).

Highlights of simulation results

- Not surprisingly, this method works well when we have matched data (matched at the same time points).
- Even when the data is not always matched, which is the primary reason to consider binning, estimates of association may be close, especially when an autoregressive process is driving the data generation, and/or when there are low levels of within-subject variability.

Highlights of simulation results

- Not surprisingly, this method works well when we have matched data (matched at the same time points).
- Even when the data is not always matched, which is the primary reason to consider binning, estimates of association may be close, especially when an autoregressive process is driving the data generation, and/or when there are low levels of within-subject variability.
- When the mismatching is extremely high, near 100%, then, only in special cases such as a high autocorrelation and/or very low levels of within-subject variability, will we see possibly acceptable levels of association bias toward the null.

Highlights of simulation results (cont.)

Highlights of simulation results (cont.)

- Bias may be the result under high mismatching, but this will not necessarily remove a detected signal — it may just provide a very conservative estimate of association.

Highlights of simulation results (cont.)

- Bias may be the result under high mismatching, but this will not necessarily remove a detected signal — it may just provide a very conservative estimate of association.
- Bin size typically had less of an effect on bias than did other factors, though larger bins did better under high mismatching.

Highlights of simulation results (cont.)

- Bias may be the result under high mismatching, but this will not necessarily remove a detected signal — it may just provide a very conservative estimate of association.
- Bin size typically had less of an effect on bias than did other factors, though larger bins did better under high mismatching.
- When all else is equal, not surprisingly the method performs better for more obs within a fixed bin size.

What have we seen

What have we seen

- For determining the association between a continuous predictor and binary (event) longitudinal response, that, in generality, are measured at different time points, a relatively simple approach is to use binning, then an adaptation of generalized linear mixed effects modeling.

What have we seen

- For determining the association between a continuous predictor and binary (event) longitudinal response, that, in generality, are measured at different time points, a relatively simple approach is to use binning, then an adaptation of generalized linear mixed effects modeling.

- Lagged associations are easily investigated and can possibly provide answers to potentially important biomedical questions.

What have we seen

- For determining the association between a continuous predictor and binary (event) longitudinal response, that, in generality, are measured at different time points, a relatively simple approach is to use binning, then an adaptation of generalized linear mixed effects modeling.

- Lagged associations are easily investigated and can possibly provide answers to potentially important biomedical questions.

- Simulation results provide some guidelines when this method could be worth using.

Future work

Future work

1. model selection/evaluation

Future work

1. model selection/evaluation
2. random effect structure and serial correlation

Future work

1. model selection/evaluation
2. random effect structure and serial correlation
3. weighting data within bins

Future work

1. model selection/evaluation
2. random effect structure and serial correlation
3. weighting data within bins
4. theoretical properties of estimators when implementing binning

Future work

1. model selection/evaluation
2. random effect structure and serial correlation
3. weighting data within bins
4. theoretical properties of estimators when implementing binning
5. dropout

Future work

1. model selection/evaluation
2. random effect structure and serial correlation
3. weighting data within bins
4. theoretical properties of estimators when implementing binning
5. dropout
6. develop curve-based approach