

Methods for Analyzing Longitudinal Health Survey Data

Theory and
Applications

Mary E. Thompson

University of Waterloo



Acknowledgments



NATIONAL PROGRAM
ON
COMPLEX DATA
STRUCTURES

Acknowledgments

The support of the following agencies is acknowledged:

- Natural Sciences and Engineering Council of Canada (NSERC)
- National Program on Complex Data Structures (NPCDS)
- Canadian Institutes of Health Research (CIHR)
- Transdisciplinary Tobacco Use Research Center (TTURC) at Roswell Park
- National Institutes of Health (NIH)

Outline

- I. Observational studies and causality
- II. Overview of models for longitudinal survey data
- III. Adapting the models and methods to complex survey data
- IV. Scan of software

Part I: Observational Studies and Causality

- The nature of causality
- Experimentation and observational studies
- Replication and the multiphasic approach
- Role of theory
- Role of temporal order

Reference

- Marini, M.M. and Singer, B. (1988) Causality in the Social Sciences. *Sociological Methodology* 18, 347-409.
- M & S survey the literature on the **ontology** and **epistemology** of causality in the social sciences



The nature of causality

- Hume, D. (writing 1739-1748): The idea of causality arises from the empirical relations of contiguity, temporal succession, and constant conjunction; regularity of association, together with a “necessary connection”.

M&S P. 350-351

The nature of causality

- Causality is directional: in some sense the “cause” must exist before the “effect”
- Causality is an “if-then” relationship

The nature of causality

- A mathematical function $y=f(x)$, where x can be **assigned** (conceptually)
- “... there would be more snow in Denver if the Rocky Mountains were lower”
- An algorithm, with output determinable from input through a sequence of instructions

The nature of causality

- A process: a chain of operations/mappings, where input may be assignable at a number of places; **intervention**; **surgery** (Pearl, 2000)
- “A number of philosophers have argued... that to provide the **connection** between cause and effect which Hume called ‘the cement of the universe’, we must analyze causal relationships ... in terms of a causal **process** that connects the two events and explains their relationship.”
M&S P.361

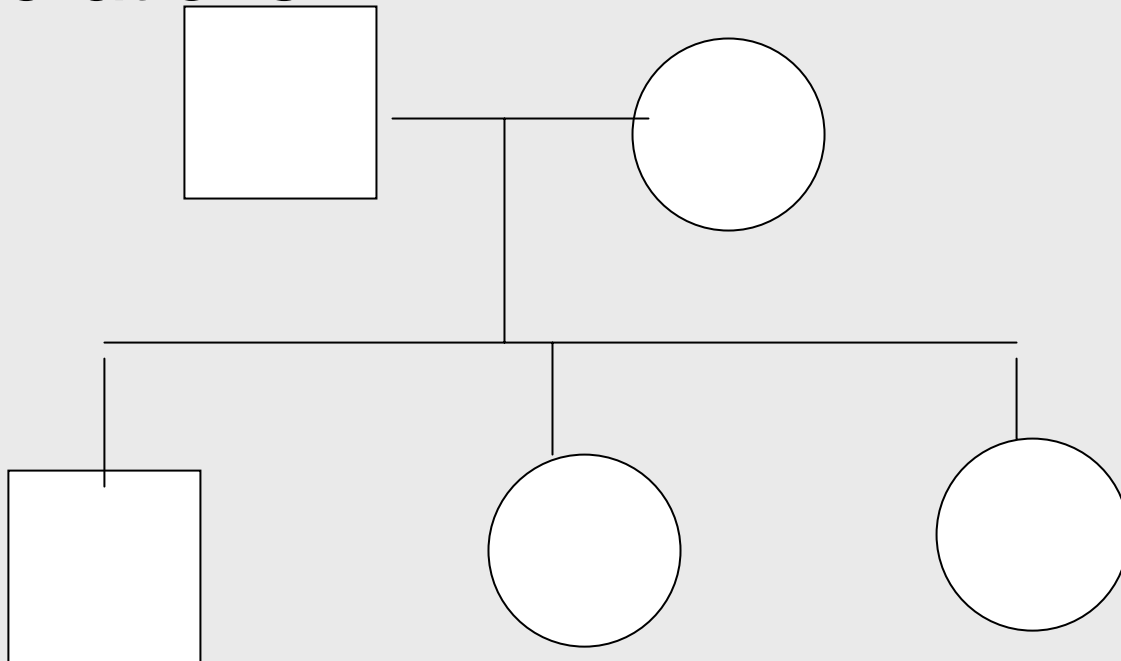
The nature of causality

Causality as information transmission:

Salmon, W.C. (1984) proposes that we view causal processes as 'the means by which structure and order are propagated or transmitted from one space-time region of the universe to other times and places'.
... An intervention at a particular point in the process transforms it in a way that persists from that point on. M&S P. 362.

The nature of causality

The life sciences show us how a stochastic element might enter the sequence of operations:



The nature of causality

- Fisher (1932, 1934): deterministic causality is “unintelligent”
- Stochasticity is the creative element in evolutionary change:
“Only in an indeterministic system has the notion of [causality] restored to it that creative element, that sense of bringing something to pass which otherwise would not have been, which is essential to its commonplace meaning.”

The nature of causality

Holders to determinism need statistical models:

“In general, the language of causation is more likely to be used when causal laws are molar, or stated in terms of large or complex objects. These laws usually involve delayed causation, **mediated via causal chains** that operate through time ...The observation of molar relationships tends to be **contingent upon many conditions**. Until these conditions are more fully known, molar causal laws will be highly fallible, and hence, probabilistic.” M&S P. 359.

The nature of causality

- When we say “smoking causes lung cancer” we mean smoking seems quite regularly to increase the chances of incurring lung cancer, and we understand something of the process or **mechanism**
- If we understood and could see the mechanism fully, we could know which smokers are destined to incur the disease, or how to prevent it

Causality in the social sciences

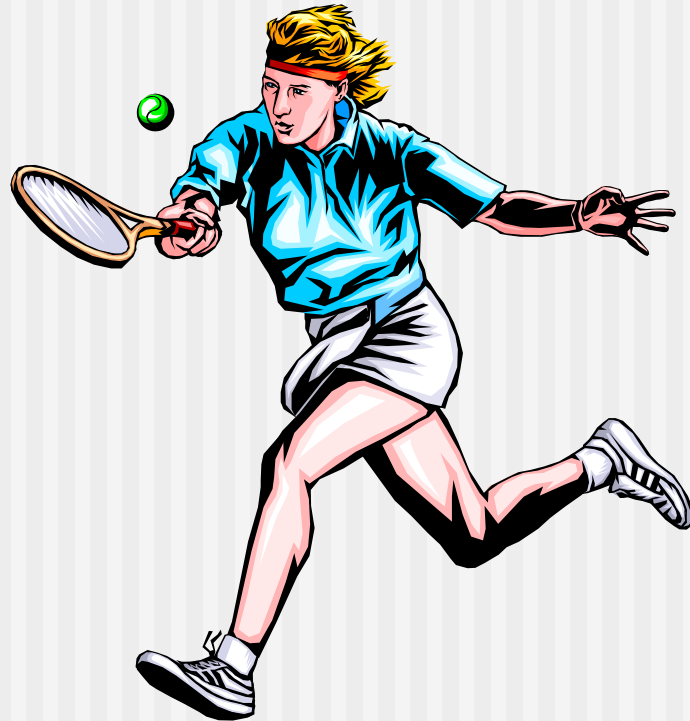
- Similar discourse
- The system (person or group) is complex, “intelligent”
- How can some factor be said to “cause” behaviour which is at least partly a matter of will?
- It is difficult to keep from influencing the processes we are trying to understand

Conditioning

- Wilson, S. E. and Howell, B. L. (2003) Do panel surveys make people sick? Arthritis trends in the Health and Retirement Study
- Shows prevalence in 55-56 age group increasing in panel cohorts (longitudinal samples) much more than in general population as indicated by NHIS; increase not accounted for by question differences or attrition bias

How do we discern a causal process?

- Everyday experimentation:



How do we discern a causal process?

- Accidental experiment:
- Landry, D. W. and Oliver, J.A. (2004) Insights into shock. *Scientific American* 290(2), 36-41.
- “ In 1997 a serendipitous observation changed the entire direction of our work.”

How do we discern a causal process?

Vasopressin restores blood pressure:

It is thought that “vasopressin reduces nitric oxide’s dilating effects on arterioles and blocks ATP-sensitive potassium channels, allowing the calcium channels to open and the [arteriole muscle] cell to contract”.

How do we discern a causal process?

- Menthol cigarettes are no more harmful than other cigarettes, but smokers of menthol cigarettes have more difficulty quitting
- Pletcher et al (2006)
- Longitudinal survey 1535 adults followed over 15 years (1200 completed)

The randomized experiment

- Allows us to discern regularity and “if-then” even under complexity, fallibility
- An association is observed between X and Y. Either
 - (i) X has caused Y or
 - (ii) Y has caused X or
 - (iii) a third variable Z has caused both.
- If the value of X has been **assigned** randomly to units, then we can rule out (ii) and (iii). Causation is established.

Observational studies

- Cannot assign treatments, at least not completely
- Cannot establish causation
- Can measure association
- Can include elements which help in understanding the association

Observational studies

- How can we understand an association?
- Can we exploit the association?
- Can we alter the association?

Replication and the multiphasic approach

- In the absence of a randomized experiment, we need a variety of (replicable) approaches.
- Example: smoking and heart disease (Doll and Hill)
- Age-adjusted mortality rates were higher for smokers than for non-smokers
- Light smokers and ex-smokers had age-adjusted mortality rates between those of non-smokers and heavy smokers.

Replication and the multiphasic approach

- Framingham heart study: 1948, random sample of 2/3 of adults aged 30 - 62 in Framingham, Massachusetts; s.s. 5209
- Physical examinations and interviews
- Offspring sample (and spouses) 5124 began in 1971
- 3500 grandchildren (study genetic influences on cardiovascular disease)

Replication and the multiphasic approach

- Rosenbaum, P. R. (2002) *Observational Studies*. 2nd edition. Springer-Verlag.
- "...when constructing a causal hypothesis one should envisage as many consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold." (Cochran, quoting Fisher)

Role of longitudinal survey

- Diggle, Liang and Zeger (1994), 1.4

$$Y_{ij} = \beta_C x_{i1} + \beta_L (x_{ij} - x_{i1}) + \varepsilon_{ij}, \quad \begin{array}{l} j = 1, \dots, n_i \\ i = 1, \dots, m \end{array}$$

- “Even when $\beta_C = \beta_L$, longitudinal studies tend to be more powerful than cross-sectional studies. The basis of inference about β_C is a comparison of individuals with a particular value of x to others with a different value”

Role of longitudinal survey

- Diggle, Liang and Zeger (1994), 1.4

$$(Y_{ij} - Y_{i1}) = \beta_L (x_{ij} - x_{i1}) + \varepsilon_{ij} - \varepsilon_{i1}$$

- “In contrast, the parameter β_L is estimated by comparing a person’s response at two [or more] times, assuming x changes with time. In a longitudinal study, each person can be thought of as serving as his or her own control.”

Role of longitudinal survey

- A temporal ordering
- A chance to observe change
- Opportunity to control for a number of possible “confounders” (Z variables)

But:

- There may be unmeasured confounders

Role of longitudinal survey

- Observations at earlier and later times are snapshots of a complex unfolding process
- Watching Batman at age 5 may predict school suspension at age 15
- Illuminating the process requires other research designs

Television and aggression

- Bushman, B. J. (1995) *J. of Personality and Social Psychology* 69, 950-960.
- Psych 101 students
- “... high trait aggressive individuals are more susceptible to the effects of violent media than are low trait aggressive individuals because they possess a relatively large network of aggressive associations that can be activated by violent cues”
- (Conjectures that prior exposure to TV violence may have built up the network....)

How can we discern causal processes?

- “To make valid causal inferences about the actions of individuals will require far more direct questioning of individuals and the mounting of longitudinal studies with successive waves of data collection spaced at short intervals.” M&S P. 401

Role of longitudinal survey

- Covariation between changes in the values of X and Y
- Possibility of combining observation with intervention
- More potential if interviews come with physical, genetic data; subject attributions
- Broad representativity (more chance of full variety of “treatments”, more chance of adjustment for confounders; more observation of rare occurrences)

Importance of attention to sampling

- In measuring association, the sampling plan defines the population to which the measurement refers
- “Achieved” sampling plan can affect the strength of the part of the association we see, e.g. in a case where non-response is higher in the upper classes and the lower classes than the middle classes.

Part II: Overview of models for longitudinal survey data

- Observational plans
- Response variables
- Explanatory variables
- Marginal models
- Random effects models
- Transition models
- Survival and event history models

Part II: Overview of models cont'd

- Missing data: censoring, omission, attrition
- Cohort and time-in-sample effects

Observational plans

- Continuous (day, hour, minute)
- Discrete, regular
- Discrete, irregular

- Single panel
- Cohorts

- Prospective, retrospective

Example: NPHS

- Begun in 1994-5 with 17,000 respondents across Canada
- Waves every two years; sample sizes
- Ideal for estimating the effects of risk factors on illness in medium and longer term
- Useful for examining precursors of risk factors

Example: NPHS

- Shields (1999) Health Reports
- No relationship between working hours and daily smoking in 1994/95, when other factors such as age and education were taken into account
- For both sexes, changing from standard to longer working hours between 1994/95 and 1996/97 was significantly associated with an increase in smoking during the period

Example: NESARC

- National Epidemiologic Survey on Alcohol and Related Conditions
- Waves in 2001-2 and 2004-5
- Wave 1 sample: 43,093 completed interviews
- To increase understanding of the natural history of alcohol use disorders and associated disabilities
- To ... identify factors that impact on their remission, chronicity, stability and initiation

Example: Dunedin Study

- 1037 babies enrolled in 1982
- Assessed ages 3, 5, 7, 9, 11,13, 15, 21,..
- 980 were assessed at age 26 in 1998-99
- Questionnaires on all aspects of lives, and physical examinations
- DNA with permission

Caspi et al (2002)

Example: Dunedin study

- Maltreatment → antisocial problems
- MAOA deficiency is a **moderator** of that relationship

Example: ITC Surveys

- International Tobacco Control Policy Evaluation Project
- Longitudinal survey of adult smokers across several countries
- Approximately annual waves
- Replenishment: new cohort each wave approximately the size of those lost to attrition

Longitudinal response variables of interest

Continuous or categorical (discrete)

Categorical: binary, ordinal, nominal, count

- Development, growth, progress
- Transition, change
- Trajectories
- Repeated measures
- Event times

Explanatory variables

- Fixed
- Time dependent
- Internal, external
- Individual, population level

Example: repeated measures

- Indonesian Children's Health Study (Diggle, Liang and Zeger 1994)
- Subsample of 275 preschoolers
- Examined quarterly for up to 6 visits
- Dependent variable presence of respiratory infection ($Y_t : t = 1, \dots, 6$)
- Main explanatory variable xerophthalmia (vitamin A deficiency)
- Estimate increase in risk of respiratory infection for children who are vitamin A deficient, controlling for other factors.

Marginal model

$$E(Y_{it}) = \mu_{it}$$

$$\log \text{it}(\mu_{it}) = \log \frac{\mu_{it}}{1 - \mu_{it}} = \log \frac{P(Y_{it} = 1)}{P(Y_{it} = 0)} = x_{it}^T \beta$$

$$\text{Var}(Y_{it}) = \mu_{it}(1 - \mu_{it})$$

$$\text{Corr}(Y_{is}, Y_{it}) = \alpha$$

Results

- Generalized estimating equation (GEE)
- Xerophthalmia increases the log odds of respiratory infection (RI) by 0.64, controlling for gender, height for age, seasonality of RI, age at entry
- Associations between Y_{t-1} and Y_t are taken into account by estimating a covariance matrix ...

Transitional model

$H_{it} = \text{history of } Y$

$$\mu_{it}^c = E(Y_{it} | H_{it})$$

$$\text{Var}(Y_{it} | H_{it}) = v_{it}^c = v(\mu_{it}^c) \phi$$

$$\text{logit}(\mu_{it}^c) = x_{it}^T \beta^{**} (H_{it})$$

Results from a transitional model

- Y_{t-1} and Y_t are strongly associated
- For cases where $Y_{t-1} = 0$, xerophthalmia increases the log odds of RI at time t by 0.78, controlling for age and season
- For cases where $Y_{t-1} = 1$, xerophthalmia increases the log odds of RI at time t by 0.88, controlling for age and season

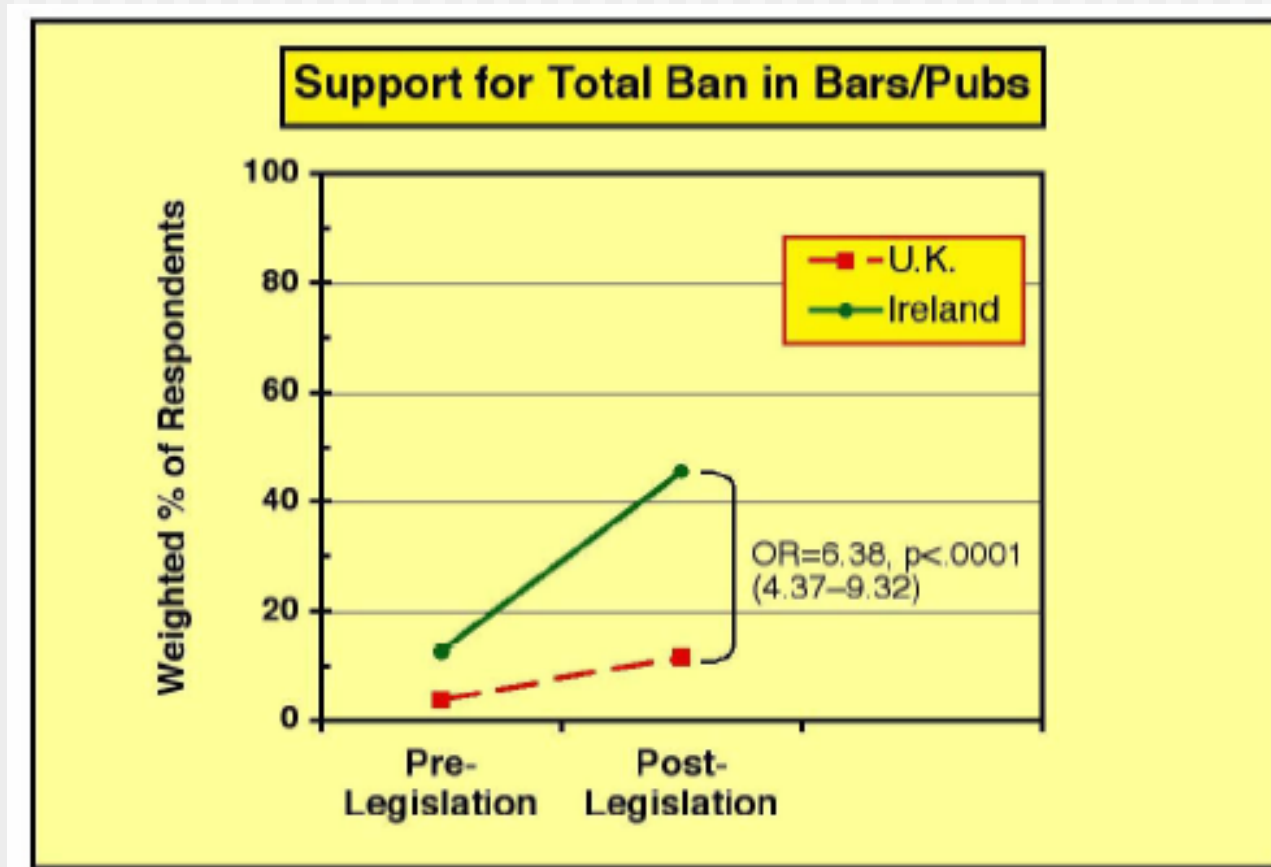
Comments

- Results (non-significant) are very similar, from marginal model, random effects model (not shown) and transitional model
- Transitional model assists causal interpretation
- The conditioned association is not affected by the fact that xerophthalmia is to start with higher among those that are frequently ill
- Results change somewhat, conditioning on Y_{t-1} and Y_{t-2}

Example: intervention impact

- International Tobacco Control (ITC) Ireland Survey
- Interviews of 755 adult smokers in Ireland, 411 in UK, before and after smoke-free law in Ireland
- $Y_t = 1$ if in favor of ban in pubs, =0 if not, time t
- x = indicator for country
- Fundamental quantities: $\pi_{ij}^x = P(Y_1 = i, Y_2 = j | x)$

Example: intervention impact



Model for intervention impact

- GEE parameterization

$$\text{logit}[P(Y_t = 1 | x)] = \alpha + \beta x + \gamma t + \delta t * x$$

$$\text{Corr}(Y_1, Y_2) = \rho$$

- 5 parameters; the key one is δ

- Advantages:

- Matches prevalence plots
- Methodology well accepted

Transition parameterization

Initial probabilities π_i^x

Transition probabilities p_{ij}^x

$$\log \text{it}(p_{01}^x) = \eta + \zeta x \quad \log \text{it}(p_{10}^x) = \mu + \nu x$$

Up to 6 parameters

Key parameter is ζ , or $\zeta - \nu$

Transition parameterization

- Advantages
- Corresponds to a probability model
- Estimation and interpretation very simple
- Incorporation of missingness is natural
- Incorporation of design information is straightforward.

Example: longitudinal mediational model

- ITC Four Country Survey
- Interviews of adult smokers in Canada, US, UK and Australia
- Longitudinal with replenishment (2000 smokers or ex-smokers) per country per wave

Example: longitudinal mediational model

Variables at wave t:

Q_t : quit intention

H_t : concern about harm to health

C_t : concern about cost

L_t : attention to warning labels

P_t : attention to price

Mediation model

Tier 1	Tier 2	Outcome
L	H	Q
P	C	
Model	$L_t L_{t-1};$	$P_t P_{t-1};$
	$H_t H_{t-1}, L_t, L_{t-1}, P_t, P_{t-1};$	$C_t C_{t-1}, L_t, L_{t-1}, P_t, P_{t-1};$
	$\bar{O} \bar{O}^{t-1}$	other variables

Mediation model

- Four waves of data, Canada and Australia
- Level of L is higher in Canada
- In both countries, marginal level of health concern is in 'steady state' $\pi_i \cong 0.65$

- Health concern transition probabilities

$$p_{01} \cong 0.63, \quad p_{10} \cong 0.21$$

- Estimated transition matrix varies with (L_{t-1}, L_t)
- $L_{t-1} = L_t = 1$ is associated with increase in p_{01} for H over two transition periods.

Example: survival models

- Time of death
- Time of menarche
- Time of school dropout
- Time of smoking cessation

Survival models

- Continuous: the “exact” time of an event is known or knowable for each subject, e.g. time of death
- Discrete: the time is best thought of as discrete, e.g. month after pregnancy

Survivor function, hazard

- Discrete case:
- $S(j)$ = probability of “surviving” beyond time j
- $h(j) = P(T=j \mid T \geq j)$ = probability of experiencing the event at j , given it is not experienced by time $j-1$

Survivor function, hazard

- Continuous case:

$$S(t) = P(T > t) = \exp\left\{-\int_0^t h(u)du\right\}$$

- Hazard function is a transition rate, e.g. life-to-death

$$h(t)dt = P(t < T \leq t + dt \mid T > t)$$

- Theory for incorporating right censoring and attrition

Example: latent mechanism

- Cirrhosis trial (Skrondal and Rabe-Hesketh 2004)
- 488 patients with liver cirrhosis
- Randomized to treatment with hormone prednisone or to placebo, indicator x
- Responses:
 - Survival time to death or censoring (lost to follow-up or alive at the end of the observation period)
 - Repeated measurements of prothrombin, a biochemical marker of liver functioning

Semiparametric model

- Proportional hazards model

$$h_j(t) = h^0(t) \exp(v_j)$$

for patient j

- Partial likelihood $PL = \prod_r \frac{\exp(v_r)}{\sum_{j \in R(t_{(r)})} \exp(v_j)}$

$R(t_{(r)})$ = the set of patients still at risk at the
r-th failure time

Cirrhosis trial

- Time at i-th measurement for patient j is t_{ij}
- Measurement model relating observed marker to latent marker $\eta_{ij}^{(2)}$:

$$y_{ij} = \beta_0 + \eta_{ij}^{(2)} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \theta)$$

- Structural model for latent marker

$$\eta_{ij}^{(2)} = \gamma_1 t_{ij} + \gamma_2 x_j + \eta_j^{(3)}, \quad \eta_j^{(3)} \sim N(0, \psi)$$

- Random intercept linear growth model

Cirrhosis trial

- Hazard model

$$\ln h_{rj} = \ln h_{rj}^0 + \lambda \eta_{rj}^{(2)} + \alpha_4 x_j$$

- Baseline hazard $\ln h^0$ a cubic polynomial in t

Cirrhosis trial

- Reduced form hazard model

$$\ln h_{rj} = \ln \bar{h}_{rj}^0 + [\lambda\gamma_2 + \alpha_4]x_j + \lambda\eta_j^{(3)}$$

where

- α_4 direct effect
- $\lambda\gamma_2$ indirect effect

Effect of latent marker on hazard: λ

Cirrhosis trial

- Results:

- Direct treatment effect:

$$\hat{\alpha}_4 = -0.18, \quad 95\% CI: (-0.42, 0.06)$$

- Indirect treatment effect (via latent marker):

$$\hat{\lambda}\hat{\gamma}_2 = 0.25, \quad 95\% CI: (0.08, 0.41)$$

- Total treatment effect:

$$\hat{\lambda}\hat{\gamma}_2 + \hat{\alpha}_4 = 0.07, \quad 95\% CI: (-0.22, 0.35)$$

- Treatment directly helpful in reducing hazard, but with negative side effect on liver functioning

Cohort and time in sample effects

- Context: survey has panels or cohorts recruited at different waves
- Repeated measures have different means and variances, depending on cohort
- Within a cohort, measures show unexplained trends

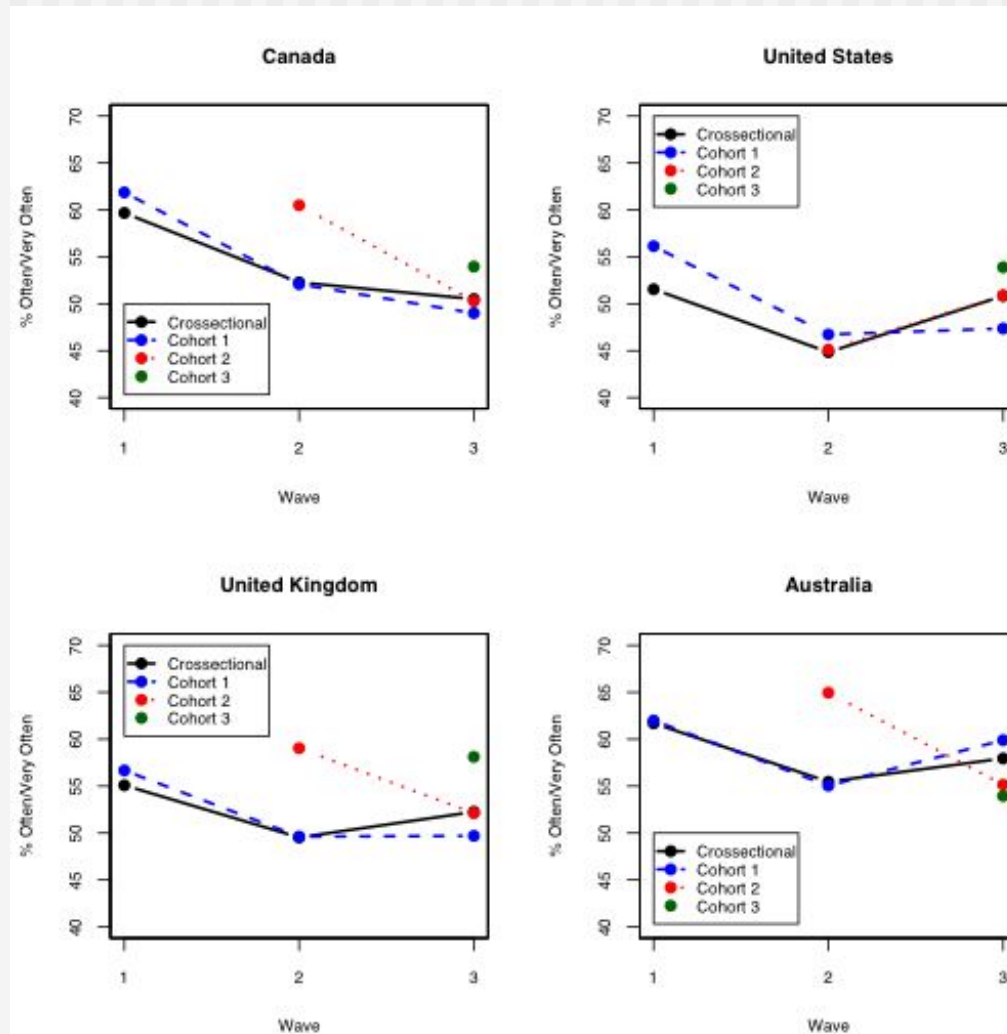
Example: ITC four country survey

Design of the ITC four-country survey

- Longitudinal within countries: Canada, U.S., U.K., Australia
- Approximately annual waves: Wave 1 in late 2002, Wave 4 in Fall 2005
- International comparisons
- RDD telephone survey
- Stratified random sampling with approximately proportional allocation
- Retention per wave has varied between 80% and 60% (lowest for U.S.)
- Sample lost is replenished with fresh cohort at each wave, with the same design as the initial one

Example: ITC 4C

Noticed Information About Dangers of Smoking



Missing data

- Item non-response
- Loss to follow-up (attrition)
- MCAR, MAR, informative missingness
- Discarding
- Imputation
- Weighting
- Modelling
- Full Information Maximum Likelihood

Part III: Adaptations to complex surveys

Features:

- Longitudinal survey designs
- Survey weights
- Effects of design on parameter interpretations
- Effects of design on precision
- Models with complex likelihoods

Example: ITC South East Asia

Sampling design in Thailand and Malaysia:

- Stratification into regions (TH), zones (MY)
- Two stage sampling of households in large urban areas
- Three stage sampling of households in rural areas
- Households sampled from 125 “clusters” in each country

Survey Weights: Definitions

initial weight

- equal to the inverse of the inclusion probability of the unit

final weight

- initial weight adjusted for nonresponse, poststratification and/or benchmarking
- interpreted as the number of units in the population that the sample unit represents

Example: ITC South-East Asia

- In Malaysia, survey weights for adults are calibrated by age, sex and ethnicity within zones
- In Thailand, survey weights for adult smokers are calibrated to assumed numbers of smokers in regions, by age and sex

Effects of design on parameter interpretations

- SRS supports iid (independent and identical distribution) assumption
- the assumption is not supported in complex surveys because of correlations induced by the sampling design or because of the population structure
- blindly applying standard programs to the analysis can lead to incorrect results

Example: ITC South-East Asia

- Models estimated with design information and without tend to be different
- We could interpret this as meaning that the corresponding parameters belong to different actual or hypothetical populations

General Effect of Complex Surveys on Precision

- stratification decreases variability (more precise than SRS)
- clustering increases variability (less precise than SRS)
- overall, the multistage design has the effect of increasing variability (less precise than SRS)

Example: ITC South-East Asia

Effects of design on precision:

- Estimates of proportions show a large design effect
- Similarly for estimates of logistic regression intercepts
- Estimates of logistic regression slopes tend to show very little design effect

Longitudinal survey design effects

Skinner and Vieira (2005)

- Design effects tend to be larger for parameters of longitudinal models than for parameters of cross-sectional models

Estimation of Variance or Precision

- variance estimation with complex multistage cluster sample design:
 - exact formula for variance estimation is often too complex; use of an approximate approach required
 - NOTE: taking account of the design in variance estimation is as crucial as using the sampling weights for the estimation of a statistic

Some Approximate Methods

- Taylor series methods
- Resampling methods
 - Balanced Repeated Replication (BRR)
 - Jackknife
 - Bootstrap

Assumptions

The resulting distribution of a test statistic is based on having a large sample size with the following properties

- the total number of first stage sampled clusters (or primary sampling units) is assumed large
 - the primary sample size in each stratum is small but the number of strata is large OR
 - the number of primary units in a stratum is large
- no survey weight is disproportionately large

Possible Violations of Assumptions

- a large-scale survey was done but inferences are desired for small subpopulations
- stratification in which a few strata (or just one) have very large sampling fractions compared to the rest of the strata
- the sampling design was uneven, resulting in large variability in the sampling weights

Resampling Methods

- the variance of an estimated parameter can be estimated by taking a large number of independent samples from the original sample
 - each new sample, called a resample, is used to estimate the parameter
 - the variability among the resulting estimates is used to estimate the variance of the full-sample estimate
 - covariance between two different parameter estimates is obtained from the covariance in replicates
- resampling methods differ in the way the resamples are built

Estimating function paradigm

Census estimating equation:

$$\sum_{i=1}^N \phi_i(y_i; \theta) = 0$$

e.g. logistic regression:

$$\sum_{i=1}^N x_i (y_i - \mu_i) = 0, \quad \mu_i = e^{x_i^T \beta} / (1 + e^{x_i^T \beta})$$

Estimating functions paradigm

Sample estimating equation:

$$\phi_s = \sum_{i \in s} w_i \phi_i = 0$$

e.g. logistic regression:

$$\sum_{i \in s} w_i x_i (y_i - \mu_i) = 0, \quad \mu_i = e^{x_i^T \beta} / (1 + e^{x_i^T \beta})$$

Estimating function paradigm

Approximate pivot:

$$\frac{\phi_s}{\sqrt{v(\phi_s)}}$$

OR

$$\hat{\theta}_s - \theta \cong -D_s^{-1} \phi_s, \quad D_s = \left(\frac{\partial \phi_s}{\partial \theta} \right)$$

$$v(\hat{\theta}_s) = \hat{D}_s^{-1} v(\phi_s) (\hat{D}_s^{-1})^T$$

Survival models

- Continuous time, proportional hazards model
- Complex survey, sampling at discrete times
- References:
 - Binder (1992) *Biometrika*
 - Lin (2000) *Biometrika*
 - Lawless and Boudreau (2006) *Canadian Journal of Statistics*
 - Rubin-Bleuer (2006)

Interaction of sampling and censoring

- Lawless (2003)
- If censoring is not informative, we condition on being present at each epoch
- Weighted Kaplan-Meier estimate assuming T and censoring time C are independent and each iid given x :
$$\hat{\lambda}(t; \tilde{x}) = \frac{\sum_{i \in s} w_i(t) d_i(t)}{\sum_{i \in s} w_i(t) y_i(t)}$$

$$w_i(t)^{-1} = \pi(x_i) \times P(C_i \geq t \mid x_i, R_i)$$

- R_i an indicator for inclusion of i in s

Ignoring unobserved heterogeneity

Muthén and Masyn (2004)

- Baseline hazard probabilities biased downward
- Time-independent covariate effects underestimated
- Spurious time-dependent effects for observed variables

Models with complex likelihoods

- Measurement error models
- Models incorporating dropout, missingness
- Partial likelihood models
- Multilevel models

Adapting the paradigm

- Replace census sums by weighted sample sums
- E.g. multilevel model:

$$\log L(\theta) = \sum_{j=1}^m w_j \log \int (\exp\{\sum_{i=1}^{n_j} w_{ij} \log f(y_{ij} | x_{ij}, \eta_j; \theta)\}) \cdot \phi(\eta_j | x_{ij}; \psi) d\eta_j$$

Weighting and replication

- Construct an artificial population by inflating the sample
- Perform the census analysis and adjust
- -- or repeatedly resample from the artificial population

Resampling methods

- Resampling can mirror the original design
- Artificial population construction and resampling can be captured in bootstrap weights

Latent transition models

- Collins, L. M. and Wugalter, S. E. (1992)
- Measurement error, measurement model
- E.g. state=quit intention classification, measured through multiple indicators (questionnaire items, etc.)
- Hidden Markov model methods
- Complex survey extension: pseudolikelihood
- Resampling to estimate standard errors
- cf Patterson, Dayton, Graubard (2002): latent class analysis in complex surveys

Example: linear growth curve analysis

- Llabre et al (2004)
- Recovery from stressors might be useful for understanding the relation between hostility and hypertension or CPD
- 167 adults aged 25-54, 74 women, 93 men
- Psychosocial variable: Cook-Medley Hostility Inventory
- Baseline systolic blood pressure (SBP)

Example: an LGC analysis

- Speech tasks (responding to wrongful accusation of shoplifting)
- Cold pressor (foot in icy water)
- SBP taken part way into tasks; 3 recovery readings 2 minutes apart
- SBP equation

$$Y_{ij} = \pi_{0j} + \pi_{1j}t_{ij} + \pi_{2j}t_{ij}^2 + r_{ij}$$

Example: an LGC analysis

- Individual coefficient dependence on hostility

$$\pi_{0j} = \beta_{00} + \beta_{01}h_j + u_{0j}$$

$$\pi_{1j} = \beta_{10} + \beta_{11}h_j + u_{1j}$$

$$\pi_{2j} = \beta_{20} + \beta_{21}h_j + u_{2j}$$

- Fit using Mplus, code provided
- Hostility “predicted” the coefficients for the speech stressor, not the cold pressor

Example: An LGC Analysis

- Multi-group analysis: allow parameters for men and women to be different, then constrain them to be equal, and test whether the fit deteriorates

Part IV: Software scan (partial)

- SUDAAN
- SAS
- SPSS
- Splus and R
- Stata, gllamm
- Mplus
- WinBUGS
- HLM, MIWin, LISREL, AMOS, ...

Checklist

- Designs accommodated
- Point estimation methods
- Variance estimation methods
- Combinability with bootstrap
- Treatment of subpopulations
- Treatment of lonely PSUs
- Treatment of missing data
- Numerical methods; accuracy; speed

Models available

- Logistic regression
- Poisson regression
- Survivor function estimation
- Proportional hazards models
- GEE
- QIF
- Mixed models; GLMM
- State space models

STATA

defining the sampling design: svyset

- example

```
svyset [pweight=indiv_wt], strata(newstrata)  
psu(ea) vce(linear)
```

- output:

```
pweight: indiv_wt  
VCE: linearized  
Strata 1: newstrata  
SU 1: ea  
FPC 1: <zero>
```

R: survey package

- define the sampling design: `svydesign`
 - `wk1de<-`
`svydesign(id=~ea,strata=~newstrata,weight=`
`~indiv_wt,nest=T,data=work1)`

- output

```
> summary(wk1de)
```

```
Stratified 1 - level Cluster Sampling design
```

```
With (1860) clusters.
```

```
svydesign(id = ~ea, strata = ~newstrata, weight = ~indiv_wt,  
nest = T, data = work1)
```

Syntax

■ STATA:

- svy: estimate
- Example: least squares estimation
- svyset [pweight=indiv_wt], strata(newstrata) psu(ea)
- svy: regress dbmi bmi

■ R:

- svy***(*, design, data=, ...)
- Example: least squares estimation
- wk2de<-
svydesign(id=~ea,strata=~newstrata,weight=~indiv_wt,nest=T,data=work2)
- svyglm(dbmi~bmi, data=work2,design=wk2de)

GEE: Generalized Estimating Equations



Dependent or response variable

- well-being measured on a 0 to 10 scale

Independent or explanatory variables'

- Has responsibility for a child under age 12 (yes = 1, no = 2)
- gender(male = 1, female = 2)
- marital status (married = 1, separated = 2, divorced = 3, never married = 5 [widowed removed from the dataset])
- employment status (employed = 1, unemployed = 2, family care = 3)

Stata syntax

```
tsset pid year, yearly
xi: xtgee wellbe i.mlstat i.job i.child i.sex
    [pweight = axrwght], family(poisson) link(identity)
    corr(exchangeable)
```

GEE Results



wellbe	Coef.	Semi-robust Std. Err.	z	P> z	[95% Conf. Interval]
_Imlstat_2	1.206905	.2036603	5.93	0.000	.8077382 1.606072
_Imlstat_3	.3732488	.120658	3.09	0.002	.1367635 .6097342
_Imlstat_5	-.0250266	.077469	-0.32	0.747	-.1768631 .1268098
_Ichild_2	-.0456858	.063007	-0.73	0.468	-.1691773 .0778056
_Ijobc_2	.9498503	.4045538	2.35	0.019	.1569394 1.742761
_Ijobc_3	.0124392	.1827747	0.07	0.946	-.3457926 .370671
_cons	1.922769	.0554797	34.66	0.000	1.814031 2.031507

For each type of marital status

Married

wellbe	Semi-robust					
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
_Ichild_2	.0666723	.0672237	0.99	0.321	-.0650836	.1984283
_Ijobc_2	.888502	.720494	1.23	0.218	-.5236403	2.300644
_Ijobc_3	.2989137	.2369747	1.26	0.207	-.1655482	.7633756
_cons	1.825918	.0562928	32.44	0.000	1.715586	1.93625

Separated or divorced

_Ichild_2	-.6732289	.1847309	-3.64	0.000	-1.035295	-.3111629
_Ijobc_2	1.239189	.8163575	1.52	0.129	-.3608422	2.83922
_Ijobc_3	-.2405778	.6582919	-0.37	0.715	-1.530806	1.049651
_cons	2.777478	.1734716	16.01	0.000	2.43748	3.117476

Never married

_Ichild_2	-.5800375	.2041848	-2.84	0.005	-.9802324	-.1798426
_Ijobc_2	.9851042	.5063179	1.95	0.052	-.0072607	1.977469
_Ijobc_3	-.2799635	.290873	-0.96	0.336	-.8500642	.2901371
_cons	2.406	.1951377	12.33	0.000	2.023538	2.788463

Random effects, categorical data, large complex surveys

Grilli and Pratesi (2004)

Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs

- SAS proc nlmixed
- Accommodation of weights through a replication option

Haynes et al (2005): HILDA survey

- 3755 women followed for 3 waves
- Employment states: ft, pt, not_e
- Time dependent explanatory variable:

$$\log\left(\frac{\pi_{itj}}{\pi_{it3}}\right) = X_{it}\beta_j + \alpha_{ij}, \quad j = 1,2$$

- gllamm (AGQ) and WinBUGS (MCMC) gave similar results, both very slow (56 hours vs 42 hours)

Other reviews

- Links from

<http://www.hcp.med.harvard.edu/statistics/survey-soft/>

- Skrondal and Rabe-Hesketh (2004)
- Chantal, K. and Suchindran, C. (2005)