2010 Australian Statistical Conference
참가 결과보고서

2010. 12.

통계청
STATISTICS KOREA

< 목　차 >

Ⅰ. ASC2010 소개 ··· 1

Ⅱ. 회의참가(출장) 개요 ··· 2

Ⅲ. 참가자 세부활동 ··· 3

<부록> : 발표자료

1. Nonparametric modeling and forecasting for electricity demand ··· 6
2. On estimation of volatility for short time series of stock prices ·· 20
3. Estimation of Population Total based on Linear Models using Social Network Information ··· 35
4. Regression Analysis Using Longitudinally Linked Data ··· 47
5. The impact of introducing CAPI to the HILDA Survey ··· 58
6. Sample Monitoring and Adjustments for Indigenous Surveys ··· 67
7. Evaluation of Feature-based Time Series Clustering ··· 75
8. A Methodology for Decomposing Age, Period and Cohort Effects Using seudo-Panel Data to Study Children's Participation in Organised Sports ·· 83
9. School segregation, class size and student achievement patterns in Salvador de Bahia (Brazil) ··· 98
10. Merging Colorectal Cancer Surveillance Data, an experience report ··· 106
11. Multivariate statistics in tax administration ··· 119
12. Sampling in the real world ··· 130
13. Trustworthy statistics - A shared responsibility? ··· 141
14. Planning for Health ··· 157
15. Weighting and Maximum Likelihood estimation to correct for errors in probabilistically linked datasets ··· 168
16. Comparing an SLK-based linkage strategy and a name-based linkage strtegy ··· 180

2010 Australian Statistical Conference 참가 결과보고

Ⅰ. ASC2010 소개

○ 표제회의(ASC2010)는 20번째 회의이며, 특히 Western Australia에서 개최되는 첫 번째 회의임

- "Statistics in the West : Understanding Our World" 라는 주제로 각분야의 전문가 및 저명한 국내외 인사들이 참가

○ 회의목적

- 통계관련 새로운 연구결과 및 진행사항을 참가자에게 정보 제공
- 통계관련 다양한 분야의 연구자간 상호 교류 협력
- 세계수준의 통계학자를 활용한 발표 및 토론을 통한 전문 지식을 공유

○ 금년회의에는 특히 10명의 저명한 기조연설자* 의 발표 및 주제별 5~6명 토론자가 참가하여 열띤 토론 진행

- Professor Barry Marshall (The University of Western Australia, winner of the 2005 Nobel Prize)
- Dr Alan M Zaslavsky (Harvard Medical School, USA)
- Professor Denise Lievesley (King's College London, UK)
- Professor Adrian Baddeley (The University of Western Australia)
- Professor Tadeusz Bednarski (Wroclaw University, Poland)
- Professor Noel Cressie (The Ohio State University, USA)
- Professor Persi Diaconis (Stanford University, USA)
- Professor Jerry Friedman (Stanford University, USA)
- Dr Gordon Smyth (WEHI, Melbourne)
- Professor Chris Wild (University of Auckland, NZ)

○ 40개 세션(통계분야)별 4~6개의 연구결과 발표 및 질의답변 진행 (총 220여개 논문 발표)

Ⅱ. 회의참가(출장) 개요

□ 목 적 :

○ 새로운 통계분석기법 도입 및 분석능력 향상을 위한 새로운 분석 프로그램 활용능력 습득

○ 선진국의 통계분야에 대한 역할, 발전 방안 등을 파악하여 향후 우리청의 다양한 분야의 통계에 대한 역할 정립을 위한 정보수집

○ 통계청의 경제통계, 사회통계, 환경통계, 조사방법론 업무담당자들의 훈련참가를 통하여 각 분야의 이론 및 실무능력을 향상시키고, 세계 각국의 통계전문가들과의 인적네트워크 형성을 통해 지속적으로 개선 개발을 도모

□ 기 간 :

○ 2010.12.4 ~ 2010.12.13 (8박 10일)

□ 장 소 :

○ Esplandae Hotel, FREMANTLE, Western Australia,

□ 출 장 자 : 총 4 명

○ 정구현(5급), 사회통계국 복지통계과

○ 강동환(6급), 경제통계국 경제기획통계과

○ 오정화(6급), 통계개발원 연구기획실

○ 안다영(7급), 통계개발원 조사연구실

Ⅲ. 참가자 세부활동

□ ASC 2010 회의 관련

※ 2010. 12.6. ~ 12.9.(4일간)

○ 회의 참관 및 자료 수집 : CD 및 책자 발간

< 회의기간중 세션별 발표논문수 >

세 션 구 분	발 표 논문(수)	비 고
1. 단독세션	**10**	
1.1. KEYNOTE ADDRESS	8	
1.2. AMSI LECTURER	1	
1.3. E K FROEMAN LECTURER	1	
2. 공동세션(통계분야별)	**220**	
2.1. Time Series(1)	5	
2.2. Time Series(2)	6	
2.3. Statistical Inference(1)	5	
2.4. Statistical Inference(2)	5	
2.5. Official Statistics	5	
2.6. CSIRO Biosciences	4	
2.7. Yong Statisticians(Where statistics can you take)	6	
2.8. Prediction	6	
2.9. Social Science	5	
2.10. Bayesian Statistics(Finance)	6	
2.11. Bayesian Statistics(Environment)	6	
2.12. Bayesian Statistics(Medicine)	6	
2.13. Mining & Multivariate Statistics	6	
2.14. Data Linkage Session	6	

2.15. Wildlife Analyses	6	
2.16. BCA Session(Current topics in Biostatistics)	4	
2.17. Survey Analysis	6	
2.18. Australian Pharmaceutical Biostatistics Group	5	
2.19. Stochastic Models	6	
2.20. Stochastic Models	5	
2.21. Stochastic Modelling	5	
2.22. Multivariate Environmental Statistics	6	
2.23. Environmental Theory	5	
2.24. Sample Design	5	
2.25. Spatial Statistics	6	
2.26. Spatial Statistics Theory	6	
2.27. Spatial/Environmental Statistics	6	
2.28. Imputation and Anslysis	6	
2.29. Inference	4	
2.30. Biostatistics(3)	5	
2.31. Biostatistics Analysis	6	
2.32. Biostatistics/Bioinformatics	6	
2.33. Biostatistics(1)	5	
2.34. Biostatistics(Design)(2)	6	
2.35. Applied Biostatistics	6	
2.36. Transport and Experimental Design	5	
2.37. Agriculture	6	
2.38. Finance	6	
2.39. Climate	6	
2.40. Environmental & Agricultural Statistics	5	

□ POST-CONFERENCE WORKSHOP

※ 2010. 12. 10. ~ 12. 11.(2일간)

○ 워크숍 참가하여 R 프로그램 활용관련 통계이론 및 실습 : 교육 실습 자료를 정리하여 CD 및 책자 발간

제목 : R 을 활용한 통계적 자료분석

1. Introduction to R	**12.10.(금)**
1.1. R 프로그램 설명	오전
1.2. R 기본명령어(Basic Syntax)	
1.3. Linear Modelling in R	
1.4. R Base Graphics	오후
1.5. Generalized Linear Models in R	
1.6. 프로그램 작성 실습	
2. Analysing spatial point patterns in R	**12.11.(토)**
2.1. Overview	오전
2.2. Data Types & Data Entry	
2.3. Intensity	
2.4. Poisson Models	
2.5. Interaction	오후
2.6. Gibbs Models	
2.7. Marked Point Patterns	
2.8. Higher Dimensions and Other Spatial Data	
2.9. 프로그램 작성실습	

1. Nonparametric modeling and forecasting for electricity demand

Nonparametric modeling and forecasting electricity demand

Han Lin Shang

Business & Economic Forecasting Unit
MONASH University

Outline

1. Problem statement
2. Data set
3. Non-dynamic updating method
4. Dynamic updating methods
5. Empirical results
6. Conclusion

Outline

1. Problem statement
2. Data set
3. Non-dynamic updating method
4. Dynamic updating methods
5. Empirical results
6. Conclusion

The problem

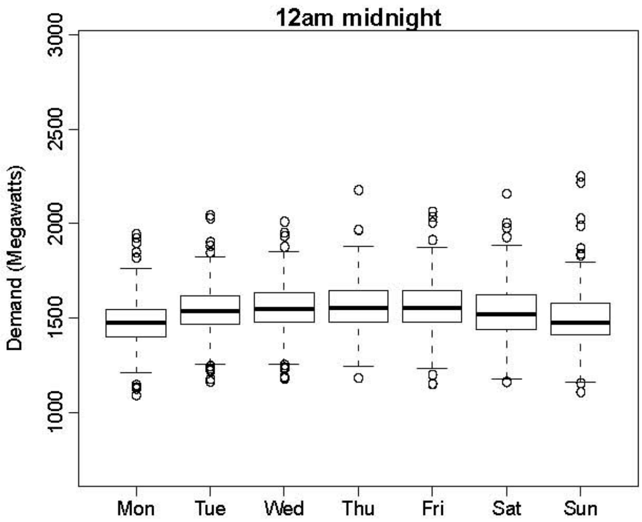
- We want to forecast the electricity demand in a half-hour period (short forecast horizon).
- We want to update the point and interval forecasts, as new observations are available.
- We have ten years of half-hourly electricity demand data (large sample size).
- The location is South Australia: home to the most volatile electricity demand in the world.

Outline

1. Problem statement
2. Data set
3. Non-dynamic updating method
4. Dynamic updating methods
5. Empirical results
6. Conclusion

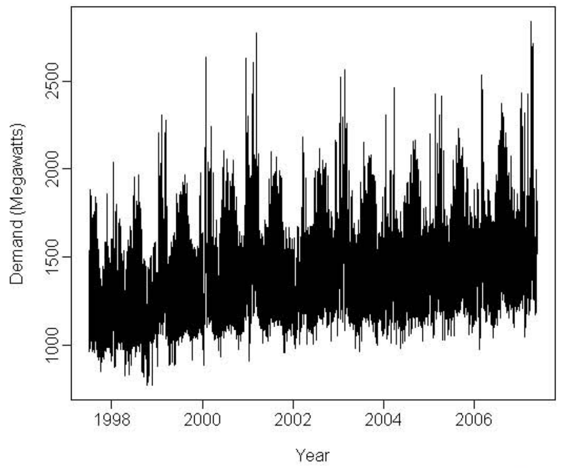
Electricity demand data and plot

- We sort electricity demand data from Mondays to Sundays.
- There is an intraday seasonality.



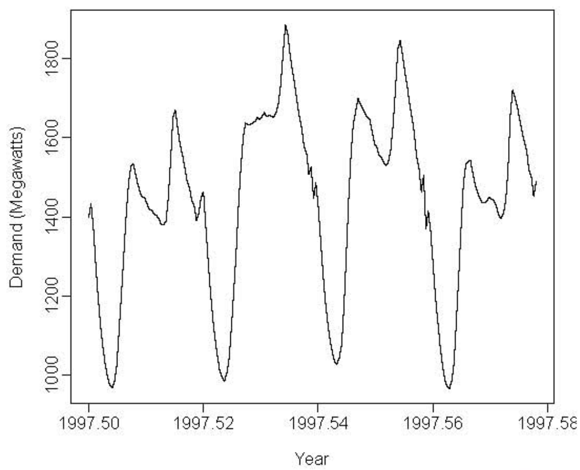
Electricity demand on Mondays

- For illustration, we display the univariate time series displays for Mondays from 7/7/1997 to 26/3/2007.
- There is an intraweek seasonality.



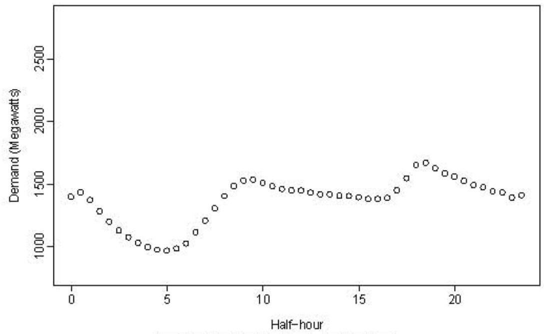
Electricity demand on Mondays

- For illustration, we display the univariate time series displays for Mondays from 7/7/1997 to 4/8/1997.
- There is an intraweek seasonality.



Functional time series display

- Slice univariate data into multivariate data.
- Reduce dimensionality from 24384 to 48.



Outline

1 Problem statement
2 Data set
3 Non-dynamic updating method
4 Dynamic updating methods
5 Empirical results
6 Conclusion

Functional time series analysis

- Let $\{Z_w, w \in [1, N]\}$ be a seasonal time series observed at N equispaced times.
- In our data set, the observed time series $\{Z_1, \dots, Z_{24384}\}$ can thus be divided into 508 trajectories of length $p = 48$.

$$y_t(x) = \{Z_w, w \in (p(t-1), pt]\},$$

where $\forall t = 1, \dots, 508$ and $\forall x \in (0, 48]$.

- The aim is to forecast future processes, denoted by $y_{n+h}(x)$; $h > 0$ from the observed data.

Principal component analysis

1. We apply principal component (pc) analysis to decompose a complete $(48 * 508)$ data matrix into a number of principal components and their associated scores.
2. The pc decomposition can be expressed as

$$y_t(x_i) = \hat{\mu}(x_i) + \sum_{k=1}^{K} \hat{\beta}_{t,k}\phi_k(x_i) + \epsilon(x_i), \qquad (1)$$

   - $\hat{\mu}(x_i)$ is the estimated time-varying mean.
   - $\hat{\beta}_{t,k}$ is the estimated time-varying principal component scores.
   - $\phi_k(x_i)$ is the principal components.
   - $\epsilon(x_i)$ is the residual term.

Forecasting method

1. Conditioning on the observed data $\mathcal{I}$ and the fixed principal components $\{\Phi = \phi_1(x_i), \dots, \phi_K(x_i)\}$, the forecast curves are expressed as

$$\hat{y}^{\text{TS}}_{n+h|n}(x_i) = \text{E}[y_{n+h}|\mathcal{I}, \Phi] = \hat{\mu}(x_i) + \sum_{k=1}^{K} \hat{\beta}_{n+h|n,k}\phi_k(x_i), \quad (2)$$

   where $\hat{\beta}_{n+h|n,k}$ denotes the h-step-ahead forecast of $\beta_{n+h,k}$ using a univariate time series forecasting method.
2. Hereafter, we call this method as the time series (TS) method.

Outline

1. Problem statement
2. Data set
3. Non-dynamic updating method
4. **Dynamic updating methods**
5. Empirical results
6. Conclusion

The problem

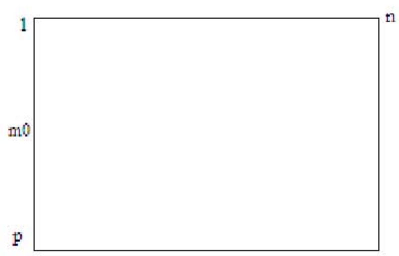
- As we observe recent data consisting of the first m_0 time period of $y_{n+1}(x_e) = [y_{n+1}(x_1), \dots, y_{n+1}(x_{m_0})]'$, we aim to update forecasts for the remaining time period of $n+1$, denoted by $y_{n+1}(x_l) = [y_{n+1}(x_{m_0+1}), \dots, y_{n+1}(x_{48})]'$.
- Using (2), the time series forecast of $y_{n+1}(x_l)$ is given as

$$\hat{y}^{\text{TS}}_{n+1|n}(x_l) = \text{E}[y_{n+1}|\boldsymbol{\mathcal{I}}_l, \boldsymbol{\Phi}_l] = \hat{\mu}(x_l) + \sum_{k=1}^{K} \hat{\beta}^{\text{TS}}_{n+1|n,k}\phi_k(x_l).$$

- TS method does not consider any new observations.
- We shall introduce four updating methods to improve point and interval forecast accuracy.

Block moving

- The block moving (BM) method considers the most recent data as the last observation in a complete data matrix.
- Because time is a continuous variable, we can observe a complete data matrix at any given time interval.
- TS method can still be applied by sacrificing a number of data points in the first year.



Ordinary least squares

1. Denote $\mathbf{F}^e$ as a $m_0 * K$ matrix whose $(j,k)^{\text{th}}$ entry is $\phi_k(x_j)$ for $1 \le j \le m_0$ and $1 \le k \le K$. Let $\boldsymbol{\beta}_{n+1} = [\beta_{n+1,1}, \dots, \beta_{n+1,K}]'$ be a $K * 1$ vector, and $\epsilon_{n+1}(x_e) = [\epsilon_{n+1}(x_1), \dots, \epsilon_{n+1}(x_{m_0})]'$ be a $m_0 * 1$ vector.
2. As the mean-adjusted $\hat{y}_{n+1}(x_e) = y_{n+1}(x_e) - \hat{\mu}(x_e)$ becomes available, we have an ordinary least squares regression expressed as
$$\hat{y}_{n+1}(x_e) = \mathbf{F}^e \boldsymbol{\beta}_{n+1} + \epsilon_{n+1}(x_e)$$
3. Via OLS, $\hat{\boldsymbol{\beta}}_{n+1} = (\mathbf{F}^{e'}\mathbf{F}^e)^{-1}\mathbf{F}^{e'}\hat{y}_{n+1}(x_e)$.
4. The OLS forecast of $y_{n+1}(x_l)$ is given by
$$\hat{y}^{\text{OLS}}_{n+1}(x_l) = \mathrm{E}[y_{n+1}(x_l)|\boldsymbol{\mathcal{I}}, \Phi(x_l)] = \hat{\mu}(x_l) + \sum_{k=1}^{K} \hat{\beta}^{\text{OLS}}_{n+1,k}\phi_k(x_l).$$

Ridge regression (RR)

1. The RR penalizes the OLS coefficients, which deviate significantly from 0. The RR coefficients minimize a penalized residual sum of squares
$$\underset{\boldsymbol{\beta}_{n+1}}{\operatorname{argmin}}\{(\hat{y}_{n+1}(x_e) - \mathbf{F}^e\boldsymbol{\beta}_{n+1})'(\hat{y}_{n+1}(x_e) - \mathbf{F}^e\boldsymbol{\beta}_{n+1}) + \lambda\boldsymbol{\beta}'_{n+1}\boldsymbol{\beta}_{n+1}\},$$
where $\lambda > 0$ is the penalty parameter.
2. By taking the derivative with respect to $\boldsymbol{\beta}_{n+1}$, we obtain
$$\hat{\boldsymbol{\beta}}^{\text{RR}}_{n+1} = (\mathbf{F}^{e'}\mathbf{F}^e + \lambda\mathbf{I})^{-1}\mathbf{F}^{e'}\hat{y}_{n+1}(x_e).$$
3. The RR forecast of $y_{n+1}(x_l)$ is given as
$$\hat{y}^{\text{RR}}_{n+1}(x_l) = \mathrm{E}[y_{n+1}(x_l)|\boldsymbol{\mathcal{I}}_l, \Phi(x_l)] = \hat{\mu}(x_l) + \sum_{k=1}^{K} \hat{\beta}^{\text{RR}}_{n+1,k}\phi_k(x_l).$$

Penalized least squares

1. The OLS method needs a sufficient number of new observations ($> K$) in order for $\hat{\beta}_{n+1}^{\text{OLS}}$ to be numerically stable.
2. β_{n+1} obtained from the PLS method minimizes

$$(\hat{y}_{n+1}(x_e) - \mathbf{F}^e\beta_{n+1})'(\hat{y}_{n+1}(x_e) - \mathbf{F}^e\beta_{n+1}) + \lambda(\hat{\beta}_{n+1} - \hat{\beta}_{n+1|n}^{\text{TS}})'(\hat{\beta}_{n+1} - \hat{\beta}_{n+1|n}^{\text{TS}})$$

3. By taking the first derivative with respect to $\hat{\beta}_{n+1}$, we obtain

$$\hat{\beta}_{n+1}^{\text{PLS}} = (\mathbf{F}^{e'}\mathbf{F}^e + \lambda\mathbf{I})^{-1}(\mathbf{F}^{e'}\hat{y}_{n+1}(x_e) + \lambda\hat{\beta}_{n+1|n}^{\text{TS}}).$$

4. The PLS forecast of $y_{n+1}(x_l)$ is given as

$$\hat{y}_{n+1}^{\text{PLS}}(x_l) = \text{E}[y_{n+1}(x_l)|\mathcal{I}_l, \Phi(x_l)] = \hat{\mu}(x_l) + \sum_{k=1}^{K}\hat{\beta}_{n+1,k}\phi_k(x_l).$$

Outline

1. Problem statement
2. Data set
3. Non-dynamic updating method
4. Dynamic updating methods
5. **Empirical results**
6. Conclusion

Penalty parameter and component selection

1. We split the data into a training set and a testing set.
2. Within the training set, we split the data into a training sample and a testing sample.
3. The optimal penalty parameter λ for different updating periods and number of principal components are determined by minimizing the mean absolute percentage errors (MAPE)

$$\text{MAPE} = \frac{1}{hp}\sum_{j=1}^{h}\sum_{i=1}^{p}\left|\frac{y_{n+j}(x_i) - \hat{y}_{n+j}(x_i)}{y_{n+j}(x_i)}\right| * 100$$

4. The optimize function in R is used to select the optimal $\lambda \in [0,1]$. The optimal number of components is the one that minimizes the MAPE.

Updating periods	RR	PLS	Updating periods	RR	PLS
1:00—23:30	0.3532	0.1378	12:30—23:30	0.4769	0.6998
1:30—23:30	0.3960	0.2038	13:00—23:30	0.4769	0.7376
2:00—23:30	0.3820	0.2566	13:30—23:30	0.6459	0.7716
2:30—23:30	0.3293	0.2469	14:00—23:30	0.6805	0.8115
3:00—23:30	0.3921	0.2548	14:30—23:30	0.3558	0.6912
3:30—23:30	0.4321	0.3240	15:00—23:30	0.3820	0.6669
4:00—23:30	0.3289	0.2016	15:30—23:30	0.3460	0.7600
4:30—23:30	0.2209	0.2196	16:00—23:30	0.2427	0.6908
5:00—23:30	0.2000	0.2409	16:30—23:30	0.1860	0.2510
5:30—23:30	0.3011	0.2574	17:00—23:30	0.3350	0.9572
6:00—23:30	0.4060	0.3620	17:30—23:30	0.0792	0.9947
6:30—23:30	0.4822	0.7108	18:00—23:30	0.0048	0.9937
7:00—23:30	0.4932	0.9874	18:30—23:30	0.0048	0.9951
7:30—23:30	0.3262	0.3475	19:00—23:30	0.4590	0.5149
8:00—23:30	0.2196	0.2196	19:30—23:30	0.2475	0.1510
8:30—23:30	0.1379	0.4080	20:00—23:30	0.3421	0.1090
9:00—23:30	0.1492	0.4438	20:30—23:30	0.3213	0.0192
9:30—23:30	0.4566	0.4803	21:00—23:30	0.0899	0.0048
10:00—23:30	0.4918	0.5147	21:30—23:30	0.0048	0.0048
10:30—23:30	0.4185	0.5573	22:00—23:30	0.0048	0.0048
11:00—23:30	0.5505	0.5917	22:30—23:30	0.0048	0.0048
11:30—23:30	0.5729	0.6262	23:00—23:30	0.0048	0.0048
12:00—23:30	0.4671	0.6656	23:30—23:30	0.0048	0.0048

Some competing methods

1. Mean predictors (MP) method predicts future observations at $t+1$ by the empirical mean from the first year to the t^{th} year.
2. Random walk (RW) method predicts future observations at year $t+1$ by the observations at year t.
3. Seasonal autoregressive moving average (SARIMA) has been considered as a benchmark method. But it requires the specifications of the order of the seasonal and non-seasonal components of an ARIMA model.
4. We implement an automatic algorithm of Hyndman and Khandakar (2008) by minimizing AIC, AICc, BIC, likelihood criteria for selecting the optimal order.

Point forecast comparison

Updating periods	Non-dynamic updating methods				Dynamic updating methods			
	MP	RW	SARIMA	TS	BM	OLS	RR	PLS
9:00—23:30	10.4246	8.7111	7.3852	7.5592	7.2023	6.2327	6.2320	6.2326
9:30—23:30	10.4118	8.7559	7.4173	7.5946	7.2633	6.2085	6.2038	6.2069
10:00—23:30	10.3977	8.7951	7.4463	7.6252	7.3626	5.9750	5.9230	5.9370
10:30—23:30	10.3835	8.8267	7.4749	7.6492	7.4509	5.7481	5.7364	5.7458
11:00—23:30	10.3714	8.8509	7.5039	7.6654	7.2585	5.6179	5.6104	5.6177
11:30—23:30	10.3626	8.8691	7.5340	7.6743	7.0264	5.5516	5.5436	5.5506
12:00—23:30	10.3513	8.8818	7.5642	7.6806	6.8234	5.6612	5.6566	5.6613
12:30—23:30	10.3443	8.8844	7.5898	7.6805	6.7780	5.7994	5.7966	5.7999
13:00—23:30	10.3476	8.8761	7.6025	7.6655	6.7226	5.9145	5.9126	5.9151
13:30—23:30	10.3529	8.8479	7.6029	7.6337	6.7328	6.0473	6.0460	6.0479
14:00—23:30	10.3515	8.8088	7.5925	7.5919	6.5597	6.2160	6.2151	6.2165
14:30—23:30	10.3562	8.7542	7.5666	7.5324	6.3209	6.3974	6.3966	6.3978
15:00—23:30	10.3648	8.6846	7.5324	7.4562	6.0053	6.5974	6.5968	6.5977
15:30—23:30	10.3733	8.5930	7.4697	7.3573	5.7507	6.7292	6.7256	6.7266
16:00—23:30	10.3760	8.4770	7.3791	7.2308	5.5620	6.4529	6.4245	6.4253
16:30—23:30	10.3695	8.3352	7.2659	7.0752	5.4975	6.0517	6.0449	6.0444
17:00—23:30	10.3544	8.1706	7.1358	6.8912	5.4994	6.2606	6.2581	6.2576
17:30—23:30	10.3336	7.9876	7.0047	6.6822	5.5173	6.5472	6.5467	6.5463
18:00—23:30	10.2881	7.7883	6.8294	6.4493	5.6643	6.9187	6.9190	6.9186
18:30—23:30	10.1663	7.5652	6.6136	6.2043	5.6307	7.3121	7.3122	7.3121
19:00—23:30	9.9810	7.3339	6.4094	6.0053	5.4354	7.3216	7.3216	7.3217
19:30—23:30	9.7757	7.1090	6.2104	5.8395	5.2703	2.9426	2.9366	2.9356
20:00—23:30	9.5759	6.8838	6.0193	5.6792	5.0626	2.5905	2.5884	2.5882
20:30—23:30	9.3964	6.6490	5.8275	5.5161	4.8095	2.2986	2.2983	2.2979
21:00—23:30	9.2094	6.3991	5.6116	5.3385	4.5848	2.0978	2.0979	2.0976
Mean	10.0832	7.8848	6.7872	6.8089	6.1855	5.5843	5.5856	5.5911

Outline

1 Problem statement
2 Data set
3 Non-dynamic updating method
4 Dynamic updating methods
5 Empirical results
6 Conclusion

Summary

1. A nonparametric forecast (TS) method is presented to forecast medium to long term electricity demand (with double seasonalities).
2. A few nonparametric updating methods are introduced for improving the point forecast accuracy. Among the updated methods of OLS, RR and PLS, there is marginal difference as measured by MAPE.
3. It may be of interest to think about other penalty functions akin to the ones in RR and PLS methods.

Paper is available at
http://robjhyndman.com/papers/dynamic-updating/

2. On estimation of volatility for short time series of stock prices

On estimation of volatility for short time series of stock prices

Nikolai Dokuchaev

Department of Mathematics & Statistics, Curtin University,

GPO Box U1987, Perth, 6845 Western Australia

email N.Dokuchaev@curtin.edu.au

November 29, 2010

Abstract

Estimation of historical volatility is considered for time series of stock prices generated by the continuous time Ito stochastic differential equations. The parameters of this equation are not assumed to be constant, their evolution law is not assumed to be known, and the frequency of data is assumed to limited. In this setting, the estimation has to be based on short time series, and the estimation error is significant. The paper suggest some supplements to the existing methods that may help to reduce the estimation error. In particular, we suggest a procedure of drift elimination via linear transformations with causal integral kernels preserving the volatility. It helps to reduce the impact of the presence of time variable and unknown drift. In addition, we suggest a modification of the standard summation formula for the volatility estimate.

Key words: econometrics, continuous time price models, discretization, short time series, volatility estimation, non-parametric estimation.

JEL classification: C14, C15, C58

Mathematical Subject Classification (2010): 91G70

1 Introduction

This short note studies estimation of historical volatility for time series of prices generated by the continuous time stochastic Ito equations. Solution of this problems is a necessary step for the volatility forecast. Once the past historical volatility is estimated, a model can be suggested for the volatility evolution law. This gives an opportunity of volatility forecast, and it is important for pricing of derivatives and optimal portfolio selection (see, e.g., [13],[15]). These problems were intensively studied; there are many well developed methods of algorithms for estimation of historical volatility and forecast of future volatilities (see, e.g., [1]-[8], [12]-[19]).

The present paper considers the problem of estimation of historical volatility. The volatility is not assumed to be constant, its evolution law is not assumed to be known, and the frequency of data is assumed to be limited. In this setting, it is unreasonable to use long-memory data for estimation. Since the volatility is not assumed to be static, older historical data are not relevant. Hence only recent observations should be used. In addition, the frequency is limited. Therefore, only short time series should be used. Because of this, the estimation error can be significant.

We suggest some modifications that may help to reduce the estimate error for volatility estimation for these models. First, we introduce a procedure of drift elimination via linear transformations with causal integral kernels preserving the volatility. It helps to reduce the impact of the presence of time variable and unknown drift (i.e., the appreciation rate of stock prices). In addition, we suggest a modification of the standard quadratic formula for volatility that uses the features of the Ito process generating the high frequency data.

2 The model

Consider a risky asset (stock, foreign currency unit, etc.) with time series of the prices $S_1, S_2, S_3, \ldots$, for example, daily prices.

The premier model of price evolution is such that $S_k = S(t_k)$, where $S(t)$ is a continuous time random process such that

$$dS(t) = S(t)[a(t)dt + \sigma(t)dw(t)].$$

Here $w(t)$ is a Wiener process, $a(t)$ is the appreciation rate, $\sigma(t)$ is the volatility, $t > 0$. We assume that a and σ are some scalar random processes such that $(a(t), \sigma(t))$ is independent from $w(\tau) - w(\theta)$ for all θ, τ such that $\theta > \tau \geq t$. We assume that the process $(a(t), \sigma(t))$ belongs to $L_2(0, T)$ with probability 1 (i.e., $\int_0^T [a(s)^2 + \sigma(s)^2]ds < +\infty$ with probability 1), for a given $T > 0$.

There are many theoretical results based on this model, including pricing of derivatives and optimal portfolio selection. Usually, practical implementation of these results requires estimation of (a, σ) from historical data. For constant a and σ, satisfactory estimates can be obtained from sufficient statistics. In fact, estimation of a for financial models is more difficult: it does not usually produce a robust result since the drift for typical financial time series is relatively small and unstable (see some references, results, and discussion in [9] and [10], Ch.9, p.128). Estimation of σ gives more robust results. This paper studies estimation of $\sigma(t)$ only.

In continuous time setting, the process $\sigma(t)$ is adapted to the filtration generated by the historical prices $S(t)$, i.e., it can be estimated without error from observation of the continuous path on the time interval $[t-\varepsilon, t]$ for an arbitrarily small $\varepsilon > 0$. In financial modelling, the continuous time model is used to describe the evolution of discrete time series of prices, and only time series of observed prices with limited frequency are available. This is the source of error in matching the statistical estimates and the continuous time model. The problem of reducing this error is studied below.

3 The estimation based on discrete time series

Let us assume that only the time series of prices $S(t_k)$ is available. We allow the volatility to be time variable, and we consider estimates at time t based on statistics collected at time $[t-\Delta t, t]$ where $\Delta t > 0$ is given.

The traditional estimate

The traditional estimate for the functional of volatility $v(t) = \frac{1}{\Delta t}\int_{t-\Delta t}^{t} \sigma(s)^2 ds$ is

$$\widehat{v}(t_m) = \frac{1}{(m-m_1)\delta} \sum_{k=m_1}^{m} (A_m - Z_k)^2, \tag{1}$$

where $t = t_m$, $t - \Delta t = t_{m_1}$, $\delta = t_k - t_{k-1}$,

$$A_m = \frac{1}{m-m_1} \sum_{k=m_1}^{m} Z_k,$$

and where

$$Z_k = \log S(t_k) - \log S(t_{k-1}).$$

(see, e.g. estimate (9.1) in [11]). We suggest some modifications of this estimate that may improve estimation. These methods are based on the assumptions that the underlying time series are generated by the model (1) and based on the properties of continuous time Ito processes.

The alternative estimate

We suggests to estimate $v(t) = \frac{1}{\Delta t}\int_{t-\Delta t}^{t}\sigma(s)^2 ds$ using the following explicit formula implied by the model (1):

$$\frac{1}{2}\int_{t-\Delta t}^{t}\sigma(s)^2 ds = \int_{t-\Delta t}^{t}\frac{dS(t)}{S(t)} - \log S(t) + \log S(t-\Delta t).$$

(See, e.g. Proposition 7.1 from [11]). It can be rewritten as

$$v(t) = \frac{2}{\Delta t}\left(\int_{t-\Delta t}^{t}\frac{dS(t)}{S(t)} - \log S(t) + \log S(t-\Delta t)\right).$$

For $t = t_m$, $t - \Delta t = t_{m_1}$, this formula leads to the following estimate of $v(t)$:

$$\widehat{v}(t_m) = \frac{2}{(m-m_1)\delta}\left(\sum_{k=m_1}^{m}\xi_k - \log S(t_m) + \log S(t_{m_1})\right), \tag{2}$$

where $\delta = t_k - t_{k-1}$,

$$\xi_k = \frac{S_k - S_{k-1}}{S_{k-1}}. \tag{3}$$

4 Reducing the impact of drift

Since only discrete $S(t_k)$ are observable, it is not possible to separate the impact of random and time variable color noise $a(t)dt$ from the impact of the noise $\sigma(t)dw(t)$.

Let $\gamma(t)$ be an adapted process, and let

$$\widehat{S}(t) = S(0) + \int_0^t \gamma(s)\widehat{S}(s)S(s)^{-1}dS(s).$$

By the definitions, $\widehat{S}(t)$ is the solution of the equation

$$d\widehat{S}(t) = \gamma(t)\widehat{S}(t)S(t)^{-1}dS(t), \qquad \widehat{S}(0) = S(0),$$

i.e.,

$$d\widehat{S}(t) = \widehat{S}(t)[\widehat{a}(t)dt + \widehat{\sigma}(t)dw(t)],$$

where

$$\widehat{a}(t) = \gamma(t)a(t), \quad \widehat{\sigma}(t) = \gamma(t)\sigma(t).$$

Lemma 4.1 *There exists a sequence of the processes $\gamma(t) = \gamma_i(t)$ such that $|\gamma_i(t)| \equiv 1$ for all i and that*

$$\int_0^T \gamma_i(t) f(t) dt \to 0 \quad \text{as} \quad i \to +\infty \quad \forall f(\cdot) \in L_2(0,T).$$

Proof. It suffices to take piecewise constant function $\gamma_i(t) = (-1)^{k(i,t)}$, where $k(i,t) = 1$ if $t \in [2mT/i, (2m+1)T/i)$, $k(i,t) = -1$ if $t \in [(2m+1)T/i, (2m+2)T/i)$, $m = 0, 1, 2, \ldots$. Clearly, the required limit holds for all $f_i \in C(0,T)$, and the set $C(0,T)$ is dense in $L_2(0,T)$. Since $\|\gamma_i\|_{L_2(0,T)} = \text{const}$, it follows that $\gamma_i \to 0$ as $i \to +\infty$ weakly in $L_2(0,T)$. This completes the proof of Lemma 4.1. $\square$

Let us consider the sequence $\{\gamma(\cdot)\} = \{\gamma_i(\cdot)\}$ from the proof of Lemma 4.1. Since

$$\widehat{S}(t) = \widehat{S}(0) + \int_0^t \gamma_i(s)\widehat{a}(s)\widehat{S}(s) ds + \int_0^t \gamma_i(s)\widehat{a}(s)\widehat{S}(s) dw(s),$$

we have that

$$\int_0^t \gamma_i(s)\widehat{a}(s)\widehat{S}(s) ds \to 0 \quad \text{a.s.},$$

and

$$\widehat{S}(t) \to \widehat{S}(0) + \int_0^t \gamma_i(s)\widehat{\sigma}(s)\widehat{S}(s) dw(s) \quad \text{as} \quad i \to +\infty \quad \text{a.s.}$$

Clearly, $\sigma(t)^2 = \widehat{\sigma}(t)^2$ and

$$\frac{1}{\Delta t}\int_{t-\Delta t}^t \sigma(s)^2 ds = \frac{1}{\Delta t}\int_{t-\Delta t}^t \widehat{\sigma}(s)^2 ds$$

Therefore, the estimate of

$$\frac{1}{\Delta t}\int_{t-\Delta t}^t \sigma(s)^2 ds$$

can be obtained via calculating the similar value for the process $\widehat{S}(t) = \widehat{S}_i(t)$ for which the impact of the appreciation rate $a(t)$ is eliminated in the limit case $i \to +\infty$.

We call $\widehat{S}(t)$ the process with eliminated drift (since $\widehat{S}(t)$ converges to a martingale).

Figure 4.1 shows an example of the simulated processes $S(t)$ and $\widehat{S}(t)$ with $\gamma(t) = \gamma_i(t)$ defined as in the proof of Lemma 4.1, with the parameters defined as

$$a(t) = 6\sin(2\pi(S(t) - 1)), \quad \sigma(t) \equiv 0.3, \quad \delta = t_k - t_{k+1} = 0.004, \tag{4}$$

where t_k are the points of discontinuity for $\gamma(t)$.

In practical calculations, the processes $\widehat{S}(t) = \widehat{S}_i(t)$ and $\gamma(t) = \gamma_i(t)$ are represented by discrete time processes.

5 The algorithm

Assume that the series of historical prices $S(t_k)$ is available, and that this is the series of data of some sufficient frequency, to justify the use of the continuous time diffusion model (1). We suggests the following procedure to estimate $v(t) = \frac{1}{\Delta t}\int_{t-\Delta t}^{t}\sigma(s)^2 ds$.

(i) Apply the drift eliminating procedure described above with $\gamma(t_k) = (-1)^k$. Let $\widehat{S}(t_k)$ be the corresponding process with eliminated drift.

(ii) Estimate the volatility using the series $\widehat{S}(t_k)$ and equation (2) or (1).

The nature of the diffusion model (1) is such that a precise enough estimate of the volatility can be achieved for the high frequency data only; decreasing the frequency leads to loss of the preciseness. Therefore, it is preferable to use the data of the highest available frequency.

Note that the drift eliminating does not take effect in a single term k under the sum in (2). However, the drift eliminating still takes effect by making the error less systematic, after mixing all $m - m_1$ terms in the sum in (2). To achieve some effect from drift elimination in a single term k in the sum in (2), the following modification of the algorithm described above can be used:

- Select $\nu \in \{1, 2, 3, ..\}$ and form the new sequence $\widehat{S}(\widehat{t}_k)$ of prices, where $\widehat{t}_k = \nu t_k$.
- Estimate the volatility using the series $\widehat{S}(\widehat{t}_k)$ and equation (2) (or, alternatively, equation (1)).

With this approach, we decrease the frequency of the series in ν times which may affect the preciseness.

6 Some experiments

Monte-Carlo simulation

In our experiments, we used Monte-Carlo simulation of the process $S(t)$ as the time series such that ξ_k in (3) is either Gaussian or has a uniform distribution, with the mean and variance selected to match the parameters of (1).

To compare different methods, we estimate the expected error

$$\mathbf{E}\left|\left(\frac{1}{\Delta t}\int_{t-\Delta t}^{t}\sigma(s)^2 ds\right)^{1/2} - \widehat{v}(t)^{1/2}\right|,$$

More precisely, we estimate the corresponding sample mean error e calculated as

$$e = \mathbb{E}\left|\left(\frac{1}{m-m_1}\sum_{k=m_1}^{m}\sigma(t_k)^2\right)^{1/2} - \widehat{v}(t_m)^{1/2}\right|,$$

and $\mathbb{E}$ denotes the sample mean over N trials in the Monte-Carlo simulation and over the number of all m. We used $N = 1,000,000$ trials with averaging over $m = 1, ..., 250$ for every Monte-Carlo trial. We found that enlarging the sample does not improve the results. Actually, the experiments with $N = 100,000$ trials produce the same results.

In the experiments, we considered only the cases of relatively small $m - m_1 \leq 10$, in the setting where we believe that volatility sustain stability only for this number of time steps.

Let us describe in detail our experiments with $\nu = 1$, with Gaussian ξ_k, and with

$$a(t) \equiv 0.3, \quad \sigma(t) \equiv 0.3, \quad t_k - t_{k+1} = 0.004, \quad m - m_1 = 10. \tag{5}$$

In the experiments with the original process with drift, estimate (1) gives the sample mean error $e = 0.0659$, and estimate (2) gives the sample mean error $e = 0.0616$. Figure 6.1 shows the corresponding estimates and the process $\sigma(t)$.

Further, let us describe the experiments with the same parameters (5) but with eliminated drift. In this case, estimate (1) gives the sample mean error $e = 0.0657$, and estimate (2) gives the sample mean error $e = 0.0616$ again. For this model with constant a, estimate (2) outperform estimate (1), and eliminating of the drift does not reduce the error. Figure 6.2 shows samples of corresponding estimates and the original process $\sigma(t)$. The error 0.0616 for estimate (2) is 7% less than the error 0.0659 for the traditional estimate (1), for the process with constant drift.

Note that the variations of parameters and equations give quite robust results, with some gradual changes of preciseness caused by changes of parameters.

Let us describe the experiments with Gaussian ξ_k, $m - m_1 = 10$, and with parameters defined by (4), i.e with time variable and random appreciation rate $a(t)$. For the original process with drift, application of estimate (1) gives the sample mean error $e = 0.0659$, and application of estimate (2) gives the sample mean error $e = 0.0826$. On the other hand, for the process with eliminated drift, application of estimate (1) gives the sample mean error $e = 0.0663$, and application of estimate (2) gives the sample mean error $e = 0.0636$. For this model, estimate (1) outperforms estimate (2) without drift eliminating. However, estimate (2) outperforms estimate (1) if the drift is eliminated, and the drift eliminating reduces the error. The error 0.0636 for estimate (2) applied with drift eliminating is 3.8% less than the error 0.0659 for the traditional estimate (i.e., when estimate (1) is used for the process with drift.

Remark 6.1 *The selection of the constant volatility in the experiments described above does not contradict to the claimed goal to study processes with time variable volatility, when long series cannot be used. In the experiments, we use short series only, i.e., the short memory of* $10 = m - m_1$ *periods. For typical sets of daily prices, it corresponds to a model such that the volatility is not supposed to stay the same for longer than a fortnight period.*

Experiment with historical prices

We have carried out some experiments for the time series representing the returns for the historical stock prices Using daily price data from 1984 to 2009 for 19 American and Australian stocks (Citibank, Coca Cola, IBM, AMC, ANZ, LEI, LLC, LLN, MAY, MLG, MMF, MWB, MIM, NAB, NBH, NCM, NCP, NFM and NPC), we generated samples of price data for one synthetic price process $S(t_k)$. In fact, the full 47 years of data was not available for all the stocks; we have the size of sample equal to 69,948.

For real prices, we don't have available the "true" volatility process in this case; we don't even know if the model (1) gives a good approximation of the price evolution. Therefore, we cannot estimate the "error" in this experiment. So far, we will demonstrate only that different estimation rules produce close enough but still different distributions of random estimates.

We estimated the value of

$$\mathbf{E}\left(\frac{1}{\Delta t}\int_{t-\Delta t}^{t}\sigma(s)^2 ds\right)^{1/2}.$$

and

$$\mathrm{Var}\left(\frac{1}{\Delta t}\int_{t-\Delta t}^{t}\sigma(s)^2 ds\right)^{1/2}.$$

More precisely, we estimate the corresponding sample mean

$$\bar{\sigma} = \mathbb{E}\left[\widehat{v}(t_m)^{1/2}\right],$$

and the corresponding sample variance

$$\mathbb{V}\mathrm{ar}\,\sigma = \mathbb{E}\left[\widehat{v}(t_m)^{1/2} - \bar{\sigma}\right]^2,$$

with estimates $\widehat{v}_k$ obtained accordingly to the different rules described above. The sample mean $\mathbb{E}$ used here represents the averaging over m and over different stocks.

For $\nu = 1$, for the process with drift, estimate (1) gives $\bar{\sigma} = 0.2449$, $\mathbb{V}\mathrm{ar}\,\sigma = 0.1505$, and estimate (2) gives $\bar{\sigma} = 0.2516$, $\mathbb{V}\mathrm{ar}\,\sigma = 0.1352$. For the process with eliminated drift, estimate (1) gives $\bar{\sigma} = 0.2446$, $\mathbb{V}\mathrm{ar}\,\sigma = 0.1406$, and estimate (2) gives $\bar{\sigma} = 0.2454$, $\mathbb{V}\mathrm{ar}\,\sigma = 0.1355$.

We already found that, for Monte-Carlo simulation of the series generated by Ito equations, different estimates with or without eliminating the drift produces different estimates for the same model. For historical prices, we observe similar situation. The difference with Monte-Carlo simulation is that we don't know actually which estimate produces a smaller error. It gives a reason to use and compare the different methods for the historical prices.

7 Discussion and conclusion

We summarize our observations as the following.

(i) In some cases (not always), drift eliminating reduces the estimation error. It may happen with estimate (2) as well as with estimate (1).

(ii) In some cases (not always), rule (2) gives less error than the mainstream rule (1). It may happen with or without drift eliminating.

The gain was modest but quite systematic and robust with respect to the changes of the parameters. For example, we observed that rule (2) gives 7% less error than the mainstream rule (1) for experiments with constant a without drift eliminating.

We have not determined yet the exact classification of models that is more appropriate for one or other method; we leave it for future research. At the moment, we can state that even the fact of the existence of some alternative estimators that reduce error in some cases is quite significant and calls to use the suggested methods as a supplement to existing methods. It can lead to improvement of preciseness of volatility estimates and, therefore, can be useful for financial applications.

The significance of the preciseness of the volatility estimation can be illustrated as the following. For instance, assume that the volatility estimate is applied in option pricing as the entrance for the Black-Scholes formula. Take, for example, call option price with the exercise time $T = 1$, the initial stock price $S(0) = 1$, the risk-free short-term rate 0.03, and with the strike price 1.2. The option price calculated for volatility $\sigma = 0.4$ is 0.1016, and the option price calculated for volatility $1.05\sigma = 0.42$ is 0.1095. This means that the 5% error for volatility estimate gives 7% error for the option price.

References

[1] Ait Sahalia, Y. and Mykland, P. (2004), Estimating diffusions with discretely and possibly randomly spaced data: A general theory. *Annals of Statistics*, 32, 2186-2222.

[2] Ait-Sahalia, Y., and Yu, J. (2009). High frequency market microstructure noise estimates and liquidity measures. *Annals of Applied Statistics* 3, pp. 422457.

[3] Andersen, T. G. and Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review* 39, pp. 885905.

[4] Andersen, T.G., Bollerslev, T., Diebold, F.X., and H. Ebens,H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics* 61, pp. 4376.

[5] Andersen, T.G., Bollerslev, T., Diebold, F.X., and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica* 71, pp. 579625

[6] Barndorff-Nielsen, O.E., Graversen S.E. and Shephard, N. (2003), Power variation & stochastic volatility: a review and some new results. *Journal of Applied Probability* 41A, 133-143.

[7] Clark, P.K. (1973). A subordinated stochastic process model with finite variance for speculative prices. *Econometrica* 41, 135155.

[8] Cvitanic,J., Liptser, R., and Rozovskii, B. (2006). A filtering approach to tracking volatility from prices observed at random times *Ann. Appl. Probab.* 16, Number 3, 1633-1652.

[9] Dokuchaev, N. (2005). Optimal solution of investment problems via linear parabolic equations generated by Kalman filter. *SIAM J. of Control and Optimization* **44**, No. 4, pp. 1239-1258.

[10] Dokuchaev N. (2002). *Dynamic portfolio strategies: quantitative methods and empirical rules for incomplete information.* Kluwer Academic Publishers, Boston.

[11] Dokuchaev N. (2007). *Mathematical finance: core theory, problems, and statistical algorithms.* Routledge.

[12] Elliott, R.J., Hunter, W.C. and Jamieson, B.M. (1998). Drift and volatility estimation in discrete time. *Jour. of Economic Dynamics & Control* 22, 209-218.

[13] Fouque, J.-P., Papanicolaou, G. and Sircar, R.. (2000) *Derivatives in Financial Markets with Stochastic Volatility.* Cambridge University Press.

[14] Frey, R. and Runggaldier, W. (2001). A nonlinear filtering approach to volatility estimation with a view towards high frequency data, (2001). *International Journal of Theoretical and Applied Finance* 4, 199-210.

[15] Hull, J., White, A. (1987). The pricing of options on assets with stochastic volatilities. *Journal of Finance* 42, 381400.

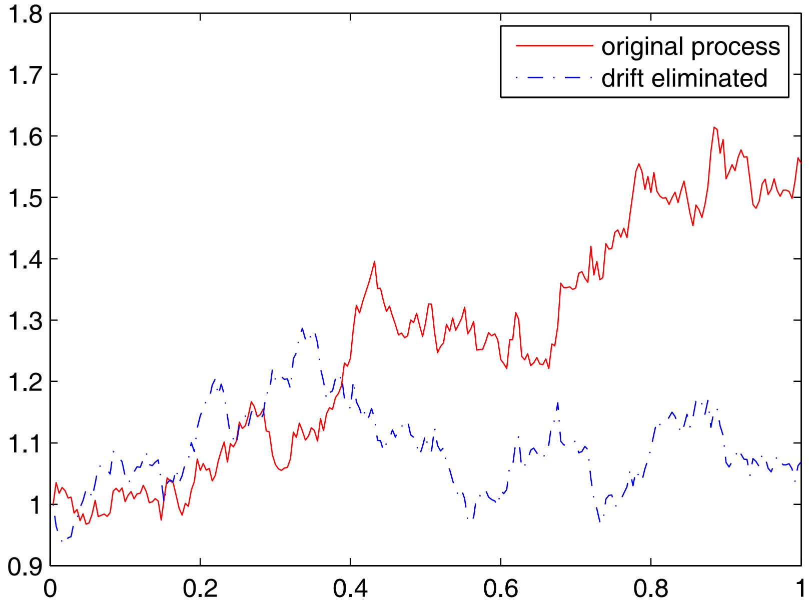
[16] Malliavin, P. and Mancino, M.E. (2002). Fourier Series method for measurement of multivariate volatilities. *Finance & Stochastics* 6, 49-62.

[17] Mandelbrot, B. (1963). The variation of certain speculative prices. *Journal of Business* 36, 394419.

[18] Merton, R.C. (1980). On estimating the expected return on the market. *Journal of Financial Economics* 8, 323361.

[19] Zhang,L., Mykland, P.A., and Ait-Sahalia, Y., A tale of two time scales: determining integrated volatility with noisy high frequency data. (2005). *Journal of the American Statistical Association* 100, pp. 13941411

Figure 4.1: Elimination of drift: $- - --$: the original process $S(t)$ with drift; $- \cdot - \cdot -\cdot$: the process $\widehat{S}(t)$ with eliminated drift. The magnified graph below demonstrates that, since the volatility dominates the appreciation rate, the difference between these two processes is barely seen from local dynamics.



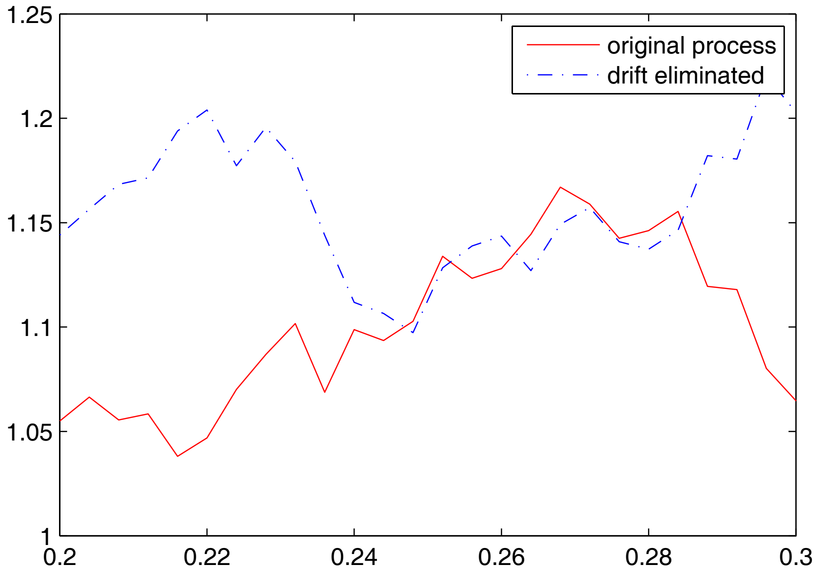


Figure 6.1: $- - -$: values of $\sigma(t)$; ——: values of $\widehat{v}(t)$ defined by (1); $- \cdot - \cdot -$: values of $\widehat{v}(t)$ defined by (2) for model (4)-(5) with drift and with $t_k - t_{k+1} = 0.004$.

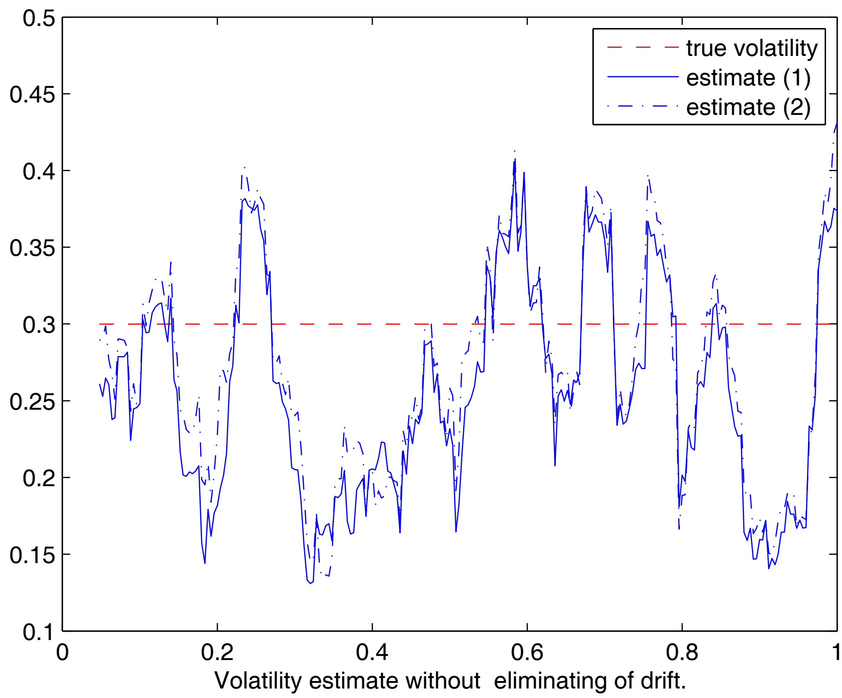
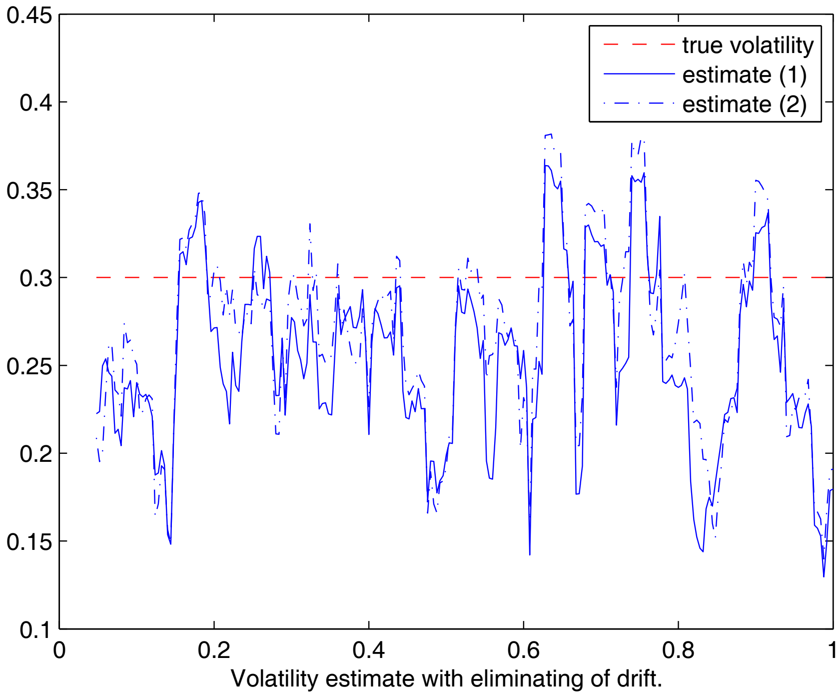


Figure 6.2: $- - -$: values of $\sigma(t)$; ——: values of $\widehat{v}(t)$ defined by (1); $- \cdot - \cdot -$: values of $\widehat{v}(t)$ defined by (2) for model (4)-(5) with eliminated drift and with $\widehat{t}_k - \widehat{t}_{k+1} = 0.002$.



3. Estimation of Population Total based on Linear Models using Social Network Information

Estimation of Population Total based on Linear Models using Social Network Information

Thomas Süße and Raymond Chambers

Centre for Statistical and Survey Methodology
University of Wollongong

ASC 2010 – Perth
Monday 6 December 2010

University of Wollongong

Outline

1. BHPS

University of Wollongong

Outline

1. BHPS
2. Social Networks

University of Wollongong

Outline BHPS Social Networks Simulation Study Conclusion

Outline

1. BHPS
2. Social Networks
3. Simulation Study

University of Wollongong

Outline

1. BHPS
2. Social Networks
3. Simulation Study
4. Conclusion

University of Wollongong

British Household Panel Study (BHPS)

- British Household Panel Study, annual longitudinal survey in UK from 1991
- Representative sample of about $5,500$ households covering more than $10,000$ individuals
- Main objective of the survey is to further our understanding of social and economic change at the individual and household level in Britain
- The BHPS provides information about surveyed individuals but also information about the three closest friends obtained from surveyed person

University of Wollongong

British Household Panel Study (BHPS)

- Information is available in seven waves: even-numbered years 1992-2004
- Collected variables of 3 best friends: sex, age, duration of friendship, frequency of contact, distance to friends, job/employment status, ethnicity
- Interested in estimating population total of annual income
- Consider model-based approach (and not model assisted and design-based)

University of Wollongong

British Household Panel Study (BHPS)

- Linear Model with p explanatory variables $\mathbf{X}_i = (X_{1i}, \ldots, X_{pi})$ and annual income Y_i

$$Y_i = \mathbf{X}_i\beta + \varepsilon_i, \varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{V})$$

- Population P of size N, sample s of size n, non-sample $r := P \setminus s$
- BLUP for estimating $T = \sum_{i \in P} Y_i$:

$$\hat{T} = \mathbf{1}_s^T\mathbf{Y}_s + \mathbf{1}_r^T\left\{\mathbf{X}_r\hat{\beta}_s + \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}(\mathbf{Y}_s - \mathbf{X}_s\hat{\beta}_s)\right\}$$

with

$$\hat{\beta}_s = (\mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\mathbf{Y}_s$$

and

$$\mathbf{Y} = \begin{pmatrix}\mathbf{Y}_s\\ \mathbf{Y}_r\end{pmatrix}, \mathbf{X} = \begin{pmatrix}\mathbf{X}_s\\ \mathbf{X}_r\end{pmatrix}\ \mathbf{V} = \begin{pmatrix}\mathbf{V}_{ss} & \mathbf{V}_{rs}\\ \mathbf{V}_{rs} & \mathbf{V}_{rr}\end{pmatrix}$$

University of Wollongong

British Household Panel Study (BHPS)

- We want to use friendship information to get more precise estimates of T
- Social relationships, such as best friends, are often represented in (social) networks
- Persons i and j are best friends, then $Z_{ij} = 1$, otherwise $Z_{ij} = 0$
- Adjacency matrix $\mathbf{Z}$ contains network information
- There are several models (auto-correlation, disturbance and contextual models) using this network information
- We also extend this idea and propose multi-level model
- BHPS only allows contextual model, because network (linking people by 3 best friends relationship) unknown

University of Wollongong

British Household Panel Study (BHPS)

- Contextual Model
$$Y_i = \mathbf{X}_i\beta + \bar{\mathbf{X}}_i\bar{\beta} + \varepsilon_i$$
- $\bar{\mathbf{X}}_i$ contains contextual variables
- $\bar{X}_i$ is average over X_j with j friend of i
- Contextual variables ($\bar{\mathbf{X}}_i$): sex and distance (indicator for < 5 miles), original variable has levels "Less than 1 mile","Less than 5 miles","5-50 miles","Over 50 miles"
- Individual level predictors ($\mathbf{X}_i$) : Age (18-24,25-34,35-44,45-54,55-64,65-74,75+), sex, highest qualification (5 categories), student (yes/no), full-time, part-time, not working, number children, household size, ~~ethnicity~~

University of Wollongong

British Household Panel Study (BHPS)

- Use wave N (2004) of BHPS as the population and select $2,000$ individuals by SRSWR (Simple random sampling without replacement)
- For simplicity we ignore household structure and spatial dependence
- Interested in estimating population total of annual income ($209,885,717$ pounds) for $14,777$ individuals for which annual income is available
- Problem: data is not normally distributed and has zeros ($\approx 4.3\%$)

University of Wollongong

British Household Panel Study (BHPS)

- Use stepwise procedure, model $\Pr(Y_i > 0)$ via logistic regression and then model $Y_i | Y_i > 0$ via linear regression
- Quasi-likelihood method (using `glm` and `geese`) with non-constant variance ($\mathrm{Var}(Y) = \phi \mathbb{E} Y$) and ML with different distributions (e.g. gamma) did not converge
- $\bar{\mathbf{X}}_r$ won't be known, because friendship information unknown for non-sample
- Could impute these variables by standard imputation methods, but base imputation on imputation methods for networks

University of Wollongong

Results for BHPS

- Results based on 2000 randomly chosen samples by SRSWR

Table: 100× Relative Mean Squared Error (relative to T^2)

	Variables		
Regression on	No Contextual	Contextual	Imputed Contextual
Y	2.152	2.071	2.073

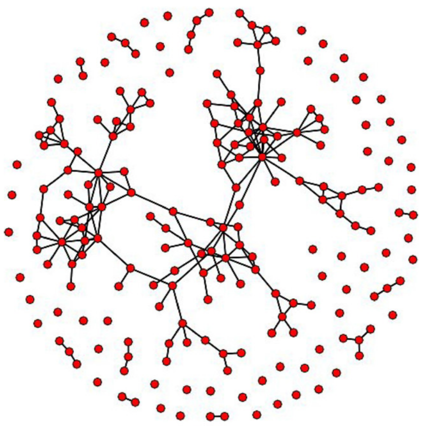
Table: 1000× Relative Bias (relative to T)

	Variables		
Regression on	No Contextual	Contextual	Imputed Contextual
Y	0.016	0.069	0.120

University of Wollongong

205 Highschool Students

- R package ”ergm” , data set ”faux.mesa.high”
- 205 high-school students in grades 7 through 12
- Links determined by mutual friendship: Each student named the other one as one of his/her top 5 friends
- Undirected Network



University of Wollongong

Exponential Random Graph Models (ERGM)

- (Curved) exponential random graph models (ERGM) are the currently most widely used models for relational data
- Probability distribution of an ERGM is of the following form

$$\Pr(\mathbf{Z}=\mathbf{z})_\theta = \exp\left(\eta(\theta)^T\mathbf{g}(\mathbf{z}) - \kappa(\eta(\theta)^T)\right) = \frac{\exp\left(\eta(\theta)^T\mathbf{g}(\mathbf{z})\right)}{\sum_{\mathbf{z}\in\mathcal{Z}}\exp(\eta(\theta)^T\mathbf{g}(\mathbf{z}))}$$

- The social network matrix $\mathbf{Z}$ can be partitioned according to the sampling process into

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_{ss} & \mathbf{Z}_{sr} \\ \mathbf{Z}_{rs} & \mathbf{Z}_{rr} \end{pmatrix}$$

- Need to impute missing networks, sampling from the conditional distribution of an ERGM, i.e. $\mathbf{Z}^{missing}|\mathbf{Z}^{observed} = \mathbf{z}^{observed}$
- Imputation methods: Conditional Independence, simple proportion approach

University of Wollongong

Simulation Study

- Consider the response variable Y_i as the income of person i and assume that education is a predictor of this income
- Let $X_i = 1,\ldots,9$ be an ordinal variable, representing education level, a low value indicates a low education level and value 9 the highest possible level of education, for example postgraduate university qualification
- We assume that the predictor for Y_i without network information is $\beta_0 + X_i\beta_1 = 40 + X_i \times 5$, which gives the total yearly income in thousands of dollars
- For example $X_i = 2$ gives an average yearly income of $40+5\times 2 = 50$ thousand Australian dollars
- Let $\sigma^2 = 1$ and $\bar{\beta} = 2.0$

University of Wollongong

Simulation Study

- Population size $N = 1,000$, sample size $n = 100$
- Just $1,000$ simulations
- Includes the situations: i) knowledge full network+model variance ii) no knowledge of network (standard linear model)
- Networks must also be generated
- ERGM: Use GWESP statistic and edges statistic, parameters 1.5 (GWESP) and -4.184591(edges)
- Artificial Network: 50 Gang networks of size 20, everybody knows everybody in gang

University of Wollongong

Simulation Results for ERGM network

Table: MSE, coverage and confidence interval (CI) length for ERGM network and contextual model

	MSE	Coverage	CI-Length
"known variance" + full network	$9,389$	96.0	395
no network information	$20,029$	96.4	601
only Z_{SS} known	$20,038$	95.6	587
Z_{SS}+Z_{SR}, cond. indep.	$10,732$	94.7	396
Z_{SS}+Z_{SR}, simple prop.	$10,379$	94.6	397

University of Wollongong

Simulation Results for Artificial Gang Network

Table: MSE, coverage and confidence interval (CI) length for artificial gang network and contextual model

	MSE	Coverage	CI-Length
”known variance” + full network	9,420	95.5	396
no network information	27,007	94.8	677
only Z_{SS} known	25,013	95.3	655
Z_{SS}+Z_{SR}	9,420	95.8	397

University of Wollongong

Conclusion

- Might be useful to collect network information to obtain higher accuracy for survey estimation
- Contextual models seem most useful compared to other 3 models (auto-Correlation, disturbance and covariance models)
- Asking people about their friends seems good option to collect contextual information instead of collecting network itself
- Contextual variables that are important predictors in certain surveys must be known
- If network is collected, then $\mathbf{Z}_{ss}$ and $\mathbf{Z}_{sr}$ plus simple imputation method seems sufficient
- Network collection has advantage, that all models can be applied and contextual information can be computed from all individual level variables
- Needs further investigation (examples and simulation studies)

University of Wollongong

Thanks

University of Wollongong

4. Regression Analysis Using Longitudinally Linked Data

Regression Analysis Using Longitudinally Linked Data

Gunky Kim
CSSM
University of Wollongong

Based on the joint work with Prof. Ray Chambers

1

Outline

- A brief review on statistical framework for linkage errors:
  - Probabilistic Record linkage
  - Record linkage and regression: Error correction models
- New development on error correction models for longitudinally linked data sets:
  - Linear Regression: Register to register case
  - Linear Regression: Sample to register case
- Futher research directions

2

Probabilistic Record Linkage

- Record linkage is a technique of linking records that refer to the same unit in two or more files.
- Applications:

  ○ Merging of large databases
  ○ Generating longitudinal records from cross-sectional data
  ○ Combining data sources

- Problem: No unique identifier in each file.
- Probabilistic Record Linkage: It links records using a set of observed variables present in both data sets (matching variables), by maximising the probability that they refer to the same unit.
- Consequence: Possible linkage errors

3

Error Correction Models: Regression Setting

- The variables $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{y}$ are the variables of interest related through a linear model

$$\mathbf{y} = \beta_0 + \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \varepsilon = \mathbf{X}\beta + \varepsilon, \quad \varepsilon \sim N(0,\sigma^2),$$

then, under the no linkage error assumption,

$$B = (\sum_q \mathbf{X}_q^T \mathbf{X}_q)^{-1} (\sum_q \mathbf{X}_q^T \mathbf{y}_q)$$

is a model-unbiased estimator of the model parameter β, since

$$E(B|\mathbf{X}) = (\sum_q \mathbf{X}_q^T \mathbf{X}_q)^{-1} \sum_q \mathbf{X}_q^T E(\mathbf{y}_q|\mathbf{X}) = \beta.$$

- It will be biased if there exist linkage errors: When the values x_{1i} are not all correctly linked with corresponding y_i and x_{2i} values, our regression model

$$\mathbf{y}_q^* = \beta_0 + \mathbf{X}_{1q}\beta_1 + \mathbf{X}_{2q}^*\beta_2^* + \varepsilon_q = \mathbf{X}_q^*\beta^* + \varepsilon_q,$$

where

$$\mathbf{y}_q^* = A_q\mathbf{y}_q, \mathbf{X}_{2q}^* = B_{2q}\mathbf{X}_{2q}$$

4

A Model for Linkage Error

$$\mathbf{y}_q^* = \mathbf{A}_q \mathbf{y}_q$$

$\mathbf{A}_q$ is an unknown random permutation matrix of order M_q, i.e. entries of $\mathbf{A}_q$ are either zero or one, with a value of one occurring just once in each row and column.

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} \quad \text{while} \quad \mathbf{y}^* = \begin{pmatrix} \mathbf{y_3} \\ \mathbf{y_2} \\ \mathbf{y_5} \\ \mathbf{y_4} \\ \mathbf{y_1} \end{pmatrix} \Rightarrow \mathbf{A} = \begin{bmatrix} 0 & 0 & \mathbf{1} & 0 & 0 \\ 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{1} \\ 0 & 0 & 0 & \mathbf{1} & 0 \\ \mathbf{1} & 0 & 0 & 0 & 0 \end{bmatrix}$$

5

Probability Record Linkage Model: Exchangeable Model (Chambers, 2008)

$$A_q = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \Rightarrow E_X(A_q) = \begin{bmatrix} \lambda_q & \gamma_q & \gamma_q & \gamma_q & \gamma_q \\ \gamma_q & \lambda_q & \gamma_q & \gamma_q & \gamma_q \\ \gamma_q & \gamma_q & \lambda_q & \gamma_q & \gamma_q \\ \gamma_q & \gamma_q & \gamma_q & \lambda_q & \gamma_q \\ \gamma_q & \gamma_q & \gamma_q & \gamma_q & \lambda_q \end{bmatrix}$$

$$\gamma_{1q} = (1-\lambda_{1q})/(M_q - 1)$$

Non-Informative Linkage Assumption

$$E_X\left(\mathbf{A}_q \mathbf{y}_q\right) = E_X\left(\mathbf{A}_q\right) E_X\left(\mathbf{y}_q\right) = \mathbf{E}_q \mathbf{X}_q \beta$$

- The linkage errors occurs at random given X

6

Case to be considered

- Possible linkage errors
  - One linkage error cases: Given a benchmark data set $\mathbf{X}_1$, linkage error can happen (between $\mathbf{X}_1$ and $\mathbf{X}_2$, not $\mathbf{y}$) or (between $\mathbf{X}_1$ and $\mathbf{y}$, not $\mathbf{X}_2$).
  - Two linkage error cases: linkage error can happen between $\mathbf{X}_1$ and $\mathbf{X}_2$ as well as between $\mathbf{X}_1$ and $\mathbf{y}$ or between $\mathbf{y}$ and $\mathbf{X}_1$ as well as between $\mathbf{y}$ and $\mathbf{X}_2$.
- Different data sets
  - All register sets
  - Sample-registers with complete linkage
  - Sample-registers with incomplete linkage.
- Probability measres
  - When λ_q are known
  - When λ_q are unknown

7

Lonitudinally linked data: All register case

- Case considering:
  - When x_{1i} is neither correctly linked with y_i, nor with x_{2i}.

$$y_q^* = \beta_0 + X_{1q}\beta_1 + X_{2q}^*\beta_2 + \varepsilon = X_q^*\beta + \varepsilon,$$

where $y_q^* = A_q y_q$, $X_{2q}^* = B_{2q} X_{2q}$.

  - X is not observable, only X^* can be observed. But, if the permutation matrix B_{2q} is known,

$$X_q = (1, X_{1q}, X_{2q}) = (1, X_{1q}, B_{2q}^T X_{2q}^*), \text{ hence}$$

$$E_{X^*}(X_q) = X_q^E = (1, X_{1q}, E_{\mathbf{B}_{2q}} X_{2q}^*).$$

By the exchangeable model assumption,

$$E_{B_{2q}} = E_{X^*}(B_{2q}) = (\lambda_{B_{2q}} - \gamma_{B_{2q}}) I_q + \gamma_{B_{2q}} 1_q 1_q^T \quad \text{and}$$

$$\lambda_{B_{2q}} = \Pr(\text{correct linkage between } X_{1q} \text{ and } X_{2q}^*)$$

$$\gamma_{B_{2q}} = \Pr(\text{incorrect linkage between } X_{1q} \text{ and } X_{2q}^*).$$

8

○ Similarly, by the exchangeable model assumption on $\mathbf{y}^*$ and non-informative linkage assumption,

$$E_{X^*}(y_q^*) = E_{X^*}(A_q y_q) = E_{X^*}(A_q) E_{X^*}(y_q) = E_{A_q} X_q^E \beta,$$

with

$$E_{A_q} = E_{X^*}(A_q) = (\lambda_{A_q} - \gamma_{A_q}) I_q + \gamma_{A_q} 1_q 1_q^T \quad \text{and}$$
$$\lambda_{A_q} = \Pr(\text{correct linkage between } X_{1q} \text{ and } y_q^*)$$
$$\gamma_{A_q} = \Pr(\text{incorrect linkage between } X_{1q} \text{ and } y_q^*).$$

9

- A ratio-type estimator: by OLS,

$$\hat{\beta}^* = \left[\sum_q (X_q^*)^T X_q^*\right]^{-1} \left[\sum_q (X_q^*)^T y_q^*\right] = \left[\sum_q (X_q^*)^T X_q^*\right]^{-1} \left[\sum_q (X_q^*)^T A_q y_q\right]$$

and

$$E_{X^*}(\hat{\beta}^*) = \left[\sum_q (X_q^*)^T X_q^*\right]^{-1} \left[\sum_q (X_q^*)^T E_{A_q} X_q^E\right] \beta = D\beta.$$

○ If the inverse of D exists with known $E_{B_{2q}}$ and E_{A_q}, a ratio-type of an unbiased estimator is of the form

$$\hat{\beta}_R = D^{-1} \hat{\beta}^*.$$

10

○ Variance of $\hat{\beta}_R$ is of the form

$$Var_{X^*}(\hat{\beta}_R) = D^{-1} Var_{X^*}(\hat{\beta}^*)(D^{-1})^T,$$

where

$$Var_{X^*}(\hat{\beta}^*) = \left[\sum_q \left(X_q^*\right)^T X_q^*\right]^{-1} \left[\sum_q \left(X_q^*\right)^T Var_{X^*}(y_q^*) X_q^*\right] \left[\sum_q \left(X_q^*\right)^T X_q^*\right]^{-1}$$

and, by the definition,

$$\begin{aligned} Var_{X^*}(y_q^*) &= E_{X^*}[Var_{X^*}(y_q^* \mid A_q)] + Var_{X^*}[E_{X^*}(y_q^* \mid A_q)] \\ &= A_q(E_{X^*}[Var_{X^*}(y_q^* \mid B_{2q})])A_q^T + A_q(Var_{X^*}[E_{X^*}(y_q^* \mid B_{2q})])A_q^T + Var_{X^*}(A_q X_q^E \beta) \\ &= Var_X(y_q) + V_C + V_{A.} \end{aligned}$$

V_A represents the variance component due to linkage errors between X_{1q} and X_{2q}, while V_C represents the variance component due to linkage errors between y_q and X_{1q}.

11

More general approach: The estimating function

- Suppose that one has $E(Y \mid \mathbf{X}) = g(\mathbf{X};\theta)$ where θ can be estimated by solving

$$\mathbf{H}(\theta) = 0,$$

and $\mathbf{H}(\theta)$ is a function that satisfies $E_X[\mathbf{H}(\theta_0)] = 0$. If $\mathbf{H}(\theta)$ is an unbiased estimating function and $\partial_\theta \mathbf{H}(\theta_0)$ is non-singular,

$$E_X[\hat{\theta} - \theta_0] \approx -[\partial_\theta \mathbf{H}(\theta_0)]^{-1} E_X[\mathbf{H}(\theta_0)] = \mathbf{0} \text{ and}$$

$$\text{var}_X(\hat{\theta}) \approx [\partial_\theta \mathbf{H}(\theta_0)]^{-1} \text{ var}_X[\mathbf{H}(\theta_0)]([\partial_\theta \mathbf{H}(\theta_0)]^{-1})^T.$$

- Our estimating function

$$\mathbf{H}(\theta) = \sum_q \mathbf{G}_q\{\mathbf{y}_q - \mathbf{f}_q\} \Longrightarrow \mathbf{H}^*(\theta^*) = \sum_q \mathbf{G}_q\{\mathbf{y}_q^* - \mathbf{f}_q^E\},$$

where $\mathbf{f}_q = E_X(\mathbf{y}_q)$ and $\mathbf{G}_q$ is a function of X_q.

- The asymptotic variance estimator is of the form

$$V_{X^*}(\hat{\theta}_3^*) = [\sum_q \mathbf{G}_q E_{A_q} \partial_\theta \mathbf{f}_q^E(\hat{\theta}_3^*)]^{-1} [\sum_q \mathbf{G}_q \hat{\Sigma}_q^{*3} \mathbf{G}_q^T] ([\sum_q \mathbf{G}_q E_{A_q} \partial_\theta \mathbf{f}_q^E(\hat{\theta}_3^*)]^{-1})^T \text{ where}$$

$$\Sigma_q^{*3} = \sigma_q^2 \mathbf{I}_q + \mathbf{V}_{C_{2q}} + \mathbf{V}_{A_q}$$

12

Simulation Results

- Model: $Y^* = 1 + 5X_1 + 8X_2^* + \varepsilon,$
- Population size: (500,500,500)
- The probability of correct linkage between $\mathbf{y}^*$ and $\mathbf{X}_1$: $\lambda_{A_q} = (1, 0.95, 0.75)$
- The probability of correct linkage between $\mathbf{X}_1$ and $\mathbf{X}_2^*$: $\lambda_{B_{2q}} = (1, 0.85, 0.8)$
- The estimators:
  - The naïve OLS estimator (ST): $\mathbf{G}_q = (X_q^*)^T$
  - The ratio-type estimator (R)
  - The Lahiri-Larsen estimator (A): $\mathbf{G}_q = (\hat{E}_{A_q} X_q^E)^T$
  - The empirical Best Linear Unbiased Estimator (C) :

$$\mathbf{G}_q = (\hat{E}_{A_q} X_q^E)^T (\hat{\sigma}_q^2 I_q + \hat{V}_{C_{2q}} + \hat{V}_{A_q})^{-1}$$

13

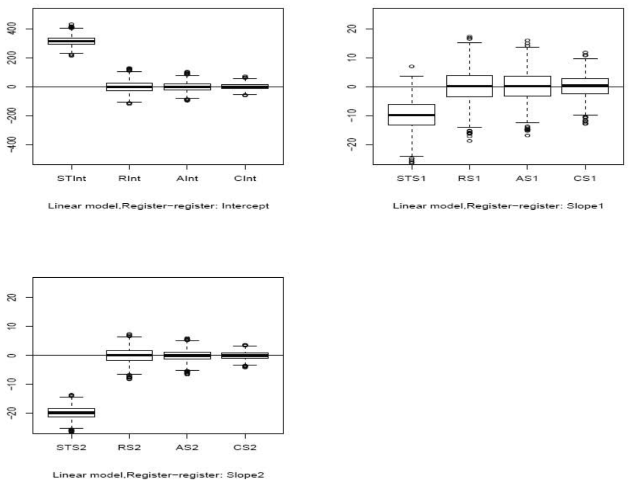


Figure 1: Simulated percentage relative errors for intercept and slope coefficients in linear regression under random linkage errors.

14

Tables

Table 1: Simulation results linear regression for register to register case: in terms of relative bias, RMSE and the actual coverage percentage for nomial 95% confidence intervals

Estimator	Relative Bias	Relative RMSE	Coverage
Simulation results for the intercept estimator			
ST	315.44	317.27	0
R	-0.61	40.18	100
A	-0.38	32.76	100
C	0.48	21.67	100
Simulation results for the first slope estimator			
ST	-9.86	24.93	50.7
R	0.14	12.55	100
A	0.13	11.34	100
C	0.08	8.73	100
Simulation results for the second slope estimator			
ST	-19.72	56.09	0
R	0.03	7.08	100
A	0.02	5.77	100
C	-0.02	3.71	100

15

Longitudinal Linkage: Sample-Register Case

- Assumption: We can choose sample s from $\mathbf{X}_1$, then link them to $\mathbf{X}_2$ and $\mathbf{y}$-registers (No non-linkage assumption:complete linkage)
- As before,

$$A_q = \begin{pmatrix} A_{sq} \\ A_{rq} \end{pmatrix} = \begin{pmatrix} A_{ssq} & A_{srq} \\ A_{rsq} & A_{rrq} \end{pmatrix}$$

which leads to

$$E_{A_q} = \begin{pmatrix} E_{A_{sq}} \\ E_{A_{rq}} \end{pmatrix} = \begin{pmatrix} E_{A_{ssq}} & E_{A_{srq}} \\ E_{A_{rsq}} & E_{A_{rrq}} \end{pmatrix}$$

and

$$\tilde{E}_{A_{sq}} = \left(\frac{\lambda_{A_q} M_q - 1}{M_q - 1}\right)\mathbf{I}_{sq} + \left(\frac{1-\lambda_{A_q}}{M_q - 1}\right)\mathbf{1}_{sq}\mathbf{w}_{sq}^T \text{ and}$$

$$\tilde{E}_{B_{2sq}} = \left(\frac{\lambda_{B_{2q}} M_q - 1}{M_q - 1}\right)\mathbf{I}_{sq} + \left(\frac{1-\lambda_{B_{2q}}}{M_q - 1}\right)\mathbf{1}_{sq}\mathbf{w}_{sq}^T.$$

16

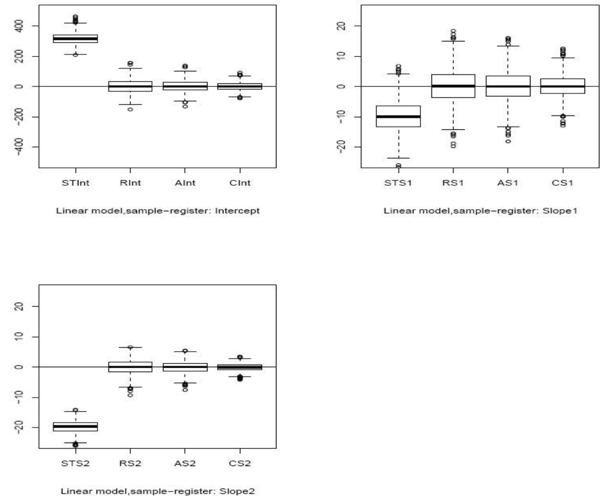


Figure 2: Simulated percentage relative errors for intercept and slope coefficients in linear regression under random linkage errors for the sample to register case.

17

Table 2: Simulation results linear regression for sample to register case: in terms of relative bias, RMSE and the actual coverage percentage for nomial 95% confidence intervals

Estimator	Relative Bias	Relative RMSE	Coverage
Simulation results for the intercept estimator			
ST	316.46	318.94	0
R	0.80	45.61	100
A	0.73	39.76	100
C	0.52	26.83	100
Simulation results for the first slope estimator			
ST	-9.96	25.24	50.1
R	0.04	12.79	100
A	0.05	11.51	100
C	0.02	8.40	100
Simulation results for the second slope estimator			
ST	-19.81	56.31	0
R	-0.10	6.86	100
A	-0.09	5.64	100
C	-0.07	3.49	100

18

Futher Research Directions

- More reasonable senarios in linkage error structure
- Loss of information issue
- Non-ignorable linkage error structure
- Dealing with small sample problem

19

References

Chambers, R. (2008). Regression analysis of probability-linked data. *Statisphere*, **4**. http://www.statisphere.govt.nz/official-statistics-research/series/vol-4.htm.

Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, **64**(328), 1183–1210.

Lahiri, P. and Larsen, M. D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, **100**(469), 222–230.

Neter, J., Maynes, E. S., and Ramanathan, R. (1965). The effect of mismatching on the measurement of response error. *Journal of the American Statistical Association*, **60**(312), 1005–1027.

Scheuren, F. and Winkler, W. E. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, **19**, 39–58.

Scheuren, F. and Winkler, W. E. (1997). Regression analysis of data files that are computer matched–part ii. *Survey Methodology*, **23**, 157–165.

20

5. The impact of introducing CAPI to the HILDA Survey

The impact of introducing CAPI to the HILDA Survey

Nicole Watson
Roger Wilkins

December 2010

THE UNIVERSITY OF MELBOURNE
FACULTY OF BUSINESS & ECONOMICS

MELBOURNE INSTITUTE® of Applied Economic and Social Research

The HILDA Survey

- Longitudinal survey of Australians with focus on family, income and labour dynamics
- Indefinite life panel with annual interviews
- Interview household
  - Household Form
  - Household Questionnaire
- Interview individuals aged 15+
  - Person Questionnaire
  - Self-Completion Questionnaire

www.melbourneinstitute.com/hilda

Why change?

- CAPI has lots to offer
  - Eliminate routing problems
  - Checking inconsistencies with respondents as they occur
  - No separate data entry
  - Improved delivery timeframes
  - Capturing paradata (eg timestamps, call record)
  - Use of dependent data
- Risk a break in the data series

www.melbourneinstitute.com/hilda

Previous experience

- Split-sample experiments
  - Longitudinal – Schrapler et al. (2006), Martin et al. (1993)
  - Cross-sectional – Fuchs et al. (2000), Lynn (1998)
  - One wave of longitudinal – Baker et al. (1995)
- Comparison over time
  - Longitudinal – Nicoletti and Peracchi (2003), Laurie (2003)
- Consistent results
  - No effect on response rates or attrition rates
  - Respondents and interviewers reacted positively
  - Routing errors by ivwrs eliminated

www.melbourneinstitute.com/hilda

Previous experience

- Mixed results
  - Refused/don’t know responses (no change vs increase)
  - Interview length (increase vs decrease)
  - Length of open ended responses (no change vs increase)
  - Social desirability bias (no change vs reducing)
  - Positioning on scale questions (no change vs more extreme)
- Most done in late ’80s and ’90s
- What may have changed?
  - Computers more widely accepted
  - Greater concern about privacy

www.melbourneinstitute.com/hilda

Dress Rehearsal split-sample test

- One wave of a longitudinal sample used to test questionnaire design and procedures
- Interviewers (rather than households) assigned to mode
  - Not random, but aimed to ensure balance between geographic spread and interviewer experience with the HILDA Survey and with computers

	Paper	CAPI
Household Questionnaire	366	343
Person Questionnaire	702	671

www.melbourneinstitute.com/hilda

Respondent characteristics

	Paper	CAPI	Diff
Male (%)	48.6	45.2	-3.4
Age (years)	42.4	44.7	2.3**
NSW (%)	51.4	50.7	-0.8
Home-owner (%)	70.2	71.6	1.4
Household size	3.2	3.1	-0.1
Marital status (%)			
Married	50.3	51.4	1.1
De facto	9.1	9.5	0.4
Separated / divorced / widow	14.5	14.9	0.4
Never married	26.1	24.0	-2.1

	Paper	CAPI	Diff
English proficiency (%)			
Speak only English	80.2	80.6	0.4
Non-Eng. speaker-English good	17.2	17.1	-0.1
Non-Eng. speaker - English poor	2.4	2.2	-0.2
Labour force status (%)			
Employed	63.5	63.6	0.1
Unemployed	4.6	3.1	-1.4
Not in the labour force	31.9	33.2	1.3

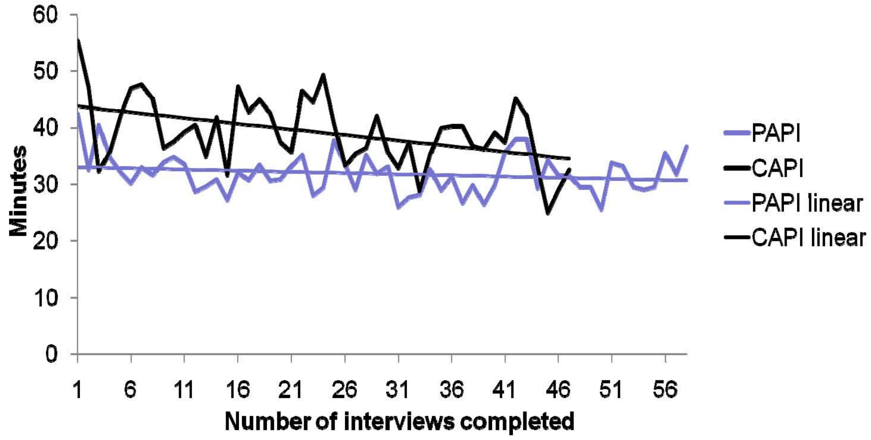
www.melbourneinstitute.com/hilda

Our findings

	Decrease	No Change	Increase
Response rates		0	
Respondent reaction		0	
Interviewer reaction		0	
Interview length			↑↑↑
Overall missingness	↓		
Don't know benefit amount			↑↑
Don't know wages/salaries amount	↓↓		
Multi-response (2 of 21 items)			↑↑
Words in open ended text			↑↑

www.melbourneinstitute.com/hilda

Interview length



www.melbourneinstitute.com/hilda

Questionnaire design issue - paper

F53a **Which ones?** [FOR EACH ONE RECEIVED, CIRCLE CORRESPONDING NUMBER IN COLUMN A BELOW.]
PROBE: **Any others? (excluding Family Tax Benefit and bonus payments already mentioned)**

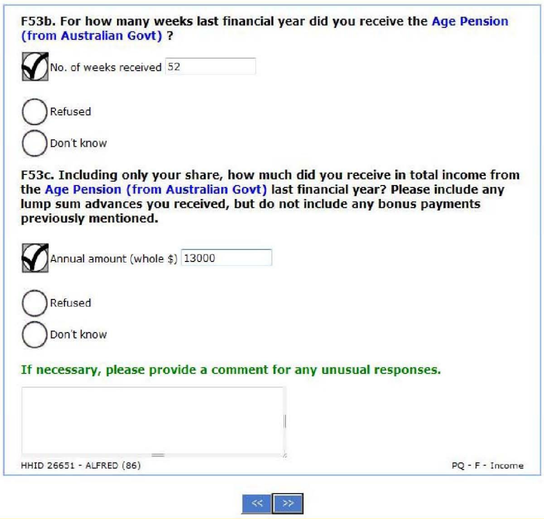
F53b **For how many weeks last financial year did you receive the** *[specify pension / allowance]***?**
[FOR EACH ONE RECEIVED, WRITE IN NUMBER IN COLUMN B BELOW.]

F53c **Including only <u>your share</u>, how much did you receive in total income from the** *[specify pension / allowance]* **last financial year? Please include any lump sum advances you received, but do not include any bonus payments previously mentioned.**
[FOR EACH ONE RECEIVED, WRITE IN AMOUNT IN COLUMN C BELOW.]
IF RESPONDENT DOES NOT KNOW YEARLY AMOUNT ASK:
What about the average received per fortnight from the *[specify pension / allowance]***? Are you able to estimate what that amount was?** WRITE IN AMOUNTS IN COLUMN D BELOW.

	A	B No. of weeks received	C Annual amount		D Average per fortnight
Age Pension (from Australian Govt)	01		$	OR	$
Newstart Allowance	02		$		$
Mature Age Allowance	03		$		$

www.melbourneinstitute.com/hilda

Questionnaire design issue - CAPI



F53b. For how many weeks last financial year did you receive the Age Pension (from Australian Govt) ?

No. of weeks received 52

Refused

Don't know

F53c. Including only your share, how much did you receive in total income from the Age Pension (from Australian Govt) last financial year? Please include any lump sum advances you received, but do not include any bonus payments previously mentioned.

Annual amount (whole $) 13000

Refused

Don't know

If necessary, please provide a comment for any unusual responses.

HHID 26651 - ALFRED (86)

PQ - F - Income

<< >>

Our findings…

	Decrease	No change	Increase
Objective data			
Labour market activity (19 items)		0	
Income (1 of 5 items)			↑↑ (benefits)
Housing (2 items)		0	
Smoking and diet (2 of 9 items)	↓↓ (vegies)		↑↑ (smoking)
Subjective data			
Satisfaction (1 of 9 items)	↓↓ (employ)		
Health (2 items)	↓ (PQ)	0	
Family and children (4 items)		0	
Reading and math skills (4 items)		0	

Introduced in 2009 (wave 9)

- Changes to CAPI
  - Used tablet and stylus
  - Integrated Household Form
- Other changes
  - New fieldwork provider
  - Increased incentive
  - New health module
- Provided cost savings
- Cannot distinguish real changes from those due to mode or change in fieldwork provider / incentive

www.melbourneinstitute.com/hilda

Findings from 2009

- Highest response rate (96.3%)
- Interview length acceptable
- Increased don't know/refused in dollar variables
- Increased amount of text in occupation / industry
- Improved consistency with use of dependent data
- Screen design important (eg annual benefit income)
- Reported values appear consistent with expectations

www.melbourneinstitute.com/hilda

Conclusions

- The 2007 Dress Rehearsal provided positive support for the move to CAPI
  - Improved data quality
    - In-field checking and editing
    - Increased length of open-ended text
    - Potentially reduces some social desirability bias
  - Concern about interview length
- Implemented successfully in 2009
  - Concern about increased missingness in dollar variables

www.melbourneinstitute.com/hilda

6. Sample Monitoring and Adjustments for Indigenous Surveys

Sample Monitoring and Adjustments for Indigenous Surveys

Tamie Anakotta
Australian Bureau of Statistics

2010

Outline

- Monitoring household surveys
- Adjusting household survey
- ABS Indigenous surveys
- Monitoring and adjustments of Indigenous surveys
- Where to from here?

Australian Bureau of Statistics

Monitoring

The ABS monitors,

- Sample Loss and Response rates
- Incompletes (not finished yet)
- Refusals

State	Fully Responding% Target = 85%	Partly Responding%	Incomplete%	Non Response%	Refusal%	Sample Loss%
AUST	85.49	1.76	0.96	9.23	2.56	13.86
ACT	86.5	3.27	0.61	6.75	2.86	9.44
NSW	82.18	1.93	0.45	12.17	3.28	14.17
NT	83.71	1.81	0.45	8.82	5.2	12.13
QLD	88.98	0.84	1.85	5.59	2.74	12.57
SA	87.71	1.14	1.2	8.3	1.65	15.25
TAS	93.38	0.41	0.14	4.14	1.93	14.5
VIC	80.98	2.97	1.55	12.46	2.04	14.45
WA	87.63	1.53	0.21	8.27	2.36	14.04

Australian Bureau of Statistics

Adjustments

Event	Impact	Adjustment
Response rates higher than expected	Increased cost	Decrease the sample size
Response rates lower than expected	Do not meet survey objectives	Increase sample size (WARNING BIAS)

ABS Surveys of Indigenous Australians

The ABS runs an Indigenous survey every 3 years, alternating between

- National Aboriginal and Torres Strait Islander Health Survey (NATSIHS)
- National Aboriginal and Torres Strait Islander Social Survey (NATSISS)

Indigenous Australian Population

- Someone who identifies themselves as being of Aboriginal origin, Torres Strait Islander origin, or both

	Indigenous Population	% of State Population
NSW	133,184	2.1%
Vic	29,143	0.6%
Qld	121,766	3.3%
SA	24,483	1.7%
WA	54,883	3.0%
Tas	16,410	3.5%
NT	51,055	27.8%
ACT	3,737	1.2%
Total	434,661	2.3%

Hit-Rate

$$hit\ rate = P(Identification) \times P(response)$$

What the ABS monitored?

- Hit-rate
- Response rates (fully and partially responding households)
- Number of fully responding adults and children
- Enumeration costs

What the ABS discovered?

- Hit rates lower than expected

State	Expected No. Indigenous HHs	Actual No. Indigenous HHs recorded
Australia	2091	1533
NSW	391	322
Vic	511	342
Qld	118	80
SA	243	172
WA	430	346
Tas	229	161
NT	88	46
ACT	81	64

Potential Causes

- Migration
- Soft-refusal
- Modal effect (face-to-face question versus paper form)
- Non-response to screening question

What the ABS did?

- Increased the sample size during enumeration
- Applied additional benchmarks to account for non-response bias

Where to from here?

- Improve collection and monitoring of screening data during enumeration
- Review screening question

Further Information

- NATSISS 08 Sample Design
  - Brent and Rogers; ABS Publication *1352.0.55.096 Sample Design Issues for National Surveys of the Indigenous Population , 2008*
- Groves, R.M., Heeringa, S.G. (2006). *Responsive design for household surveys: tools for actively controlling survey errors and costs.* Journal of Royal Statistical Society A, 169, Part 3, pp. 429-457
- Contact details

  tamie.anakotta@abs.gov.au

  +61 2 6252 7360

7. Evaluation of Feature-based Time Series Clustering

Presented by *Shen LIU*
Department of Econometrics and Business Statistics

Evaluation of Feature-based Time Series Clustering

MONASH University
Business and Economics
www.buseco.monash.edu

Why this topic?

- Why we need clustering…
  - More efficient analysis, better forecasts, and
  - *Save time and money!*
- Why feature-based…
  - Avoid extremely high dimensional input
  - Describe the dynamics of the time series efficiently
  - Enable the comparison of the time series with different lengths

MONASH University
Business and Economics

www.**buseco**.monash.edu

What has been done so far?

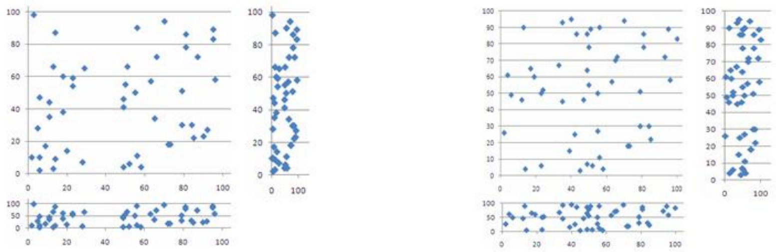
- *Liao (2005): A survey*
  - Clustering algorithms or procedures
  - Similarity/distance measures
  - Clustering results evaluation criteria
- *Parametric approaches*
  - Maharaj (2000): A clustering technique based on the p-value of the data generating processes hypotheses test
- *Non-parametric approaches*
  - Caiado et al. (2006): A new distance measure based on the normalized periodograms, which is a typical *feature-based* method
- ***Question 1***: *Which clustering method is the most effective one? Which feature of the time series tends to result in the best clustering performance?*

MONASH University
Business and Economics

www.**buseco**.monash.edu

Any drawbacks?

- The majority of the literature assigned equal weights to the time series feature values
  - *Often this is not desirable in applied works*
- ***Illustration: 40 two-dimensional time series are generated***



- ***Question 2***: *How to assign the weights to achieve better clustering performances?*

MONASH University
Business and Economics

www.buseco.monash.edu

Q1: Which method/feature is the best?

- ***2 non-hierarchical clustering methods***:
  - *k*-means: partitioning around centroids
  - *k*-medoids: partitioning around medoids (representative objects)
- ***4 hierarchical clustering methods***:
  - Single linkage: shortest distance
  - Complete linkage: maximum distance
  - Average linkage: average similarity of all individual time series in one cluster with all individual time series in another
  - Ward's method: sum of the squares within the clusters summed over all variables
- ***5 time series features:*** autocorrelation function (ACF), partial autocorrelation function (PACF), normalized periodogram (NP), log-normalized periodogram (LNP), and the cepstrum (CEP)

MONASH University
Business and Economics

www.buseco.monash.edu

Simulation design

- 15 time series from each of two autoregressive processes of order one with different parameter values, $Y_{1,t} = \phi_1 Y_{1,t-1} + a_{1,t}$ and $Y_{2,t} = \phi_2 Y_{2,t-1} + a_{2,t}$
- ϕ_1 : uniformly distributed in the range (0.3 ± 0.01)
- ϕ_2 : uniformly distributed in four ranges: (0.4 ± 0.01), (0.45 ± 0.01), (0.5 ± 0.01), (0.55 ± 0.01)
- Series length: $T = 2^n$, $n = 6,7,8,9$
- Number of the features $p = 2^2, 2^3, \ldots, 2^{n-1}$
- 1000 simulation replications
- Gaussian errors with mean zero and variance of one
- Rand Index is calculated as the cluster similarity measure

$$R = \frac{n_{SS} + n_{DD}}{n_{SS} + n_{DD} + n_{SD} + n_{DS}}$$

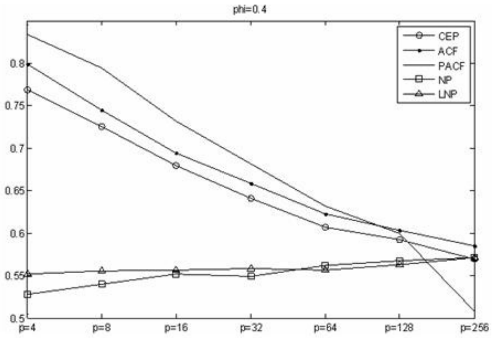
  - The greater the value, the higher the agreement between the clusters in the data set and the clusters generated by a clustering algorithm

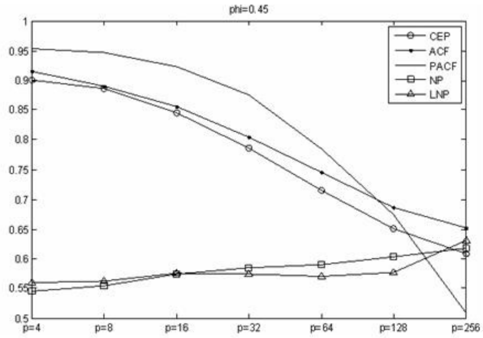
MONASH University
Business and Economics

www.buseco.monash.edu

Simulation results

- A part of the output





- Basic findings
  - When using *k*-means algorithm, the PACF feature achieves the best clustering performance

MONASH University
Business and Economics

www.buseco.monash.edu

Q2: How to assign the weights?

- **The newly proposed weighting method**
  - Step 1: For *n* time series, *calculate the feature values for p lags,* denoted by a $p \times n$ matrix $F_{p\times n}$. For the k^{th} lag ($k = 1, 2, \ldots, p$) of the feature values, evaluate its individual clustering performance by using Rand Index, denoted by R_k
  - Step 2: The weight of the k^{th} lag is calculated as

$$W_k = \frac{R_k}{\sum_{i=1}^{p} R_i}$$

  - Step 3: Multiply the k^{th} row of the matrix $F_{p\times n}$ by W_k , and the weighted matrix is denoted by $WF_{p\times n}$. Then use $WF_{p\times n}$ to cluster the time series

MONASH University
Business and Economics

www.buseco.monash.edu

Simulation results of the weighting method

- We apply this method to the best clustering combination
  - PACF features in *k*-means algorithm
- A summary table

Table 1 Comparison of the weighted results to the unweighted ones, $\phi = 0.4$

		$T = 64$	$T = 128$	$T = 256$	$T = 512$
$p = 4$	Unweighted features	0.5849	0.6364	0.7157	0.8323
	Weighted features	0.6175	0.6915	0.7805	0.8704
	Improvement	**0.0326**	**0.0551**	**0.0648**	**0.0381**
$p = 8$	Unweighted features	0.5979	0.6076	0.6746	0.7993
	Weighted features	0.6331	0.6706	0.7715	0.8659
	Improvement	**0.0352**	**0.0630**	**0.0968**	**0.0666**

Note: other *p* values have also been tried

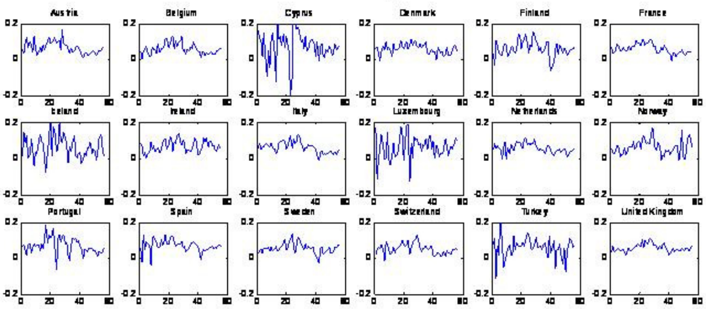
- Conclusion
  - For all series lengths and *p* values, the weighted PACF features consistently achieve better clustering performances than the unweighted ones

MONASH University
Business and Economics

www.buseco.monash.edu

Application

- Data: Annual real GDP per capita of 18 European countries
  - Austria, Belgium, Cyprus, Denmark, Finland, France, Iceland, Ireland, Italy, Luxembourg, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, Turkey, and United Kingdom
- Observation period: from 1951 to 2007 (57 observations)
- Transformed in differences of the logarithm



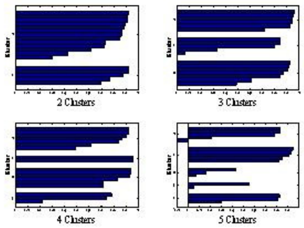
MONASH University
Business and Economics

www.**buseco**.monash.edu

Approach and result

- Use Average Silhouette coefficient (ASC) and Silhouette plot to determine the number of the clusters in data

Number of Clusters	2	3	4	5
ASC	0.7833	0.7717	**0.7945**	0.5663



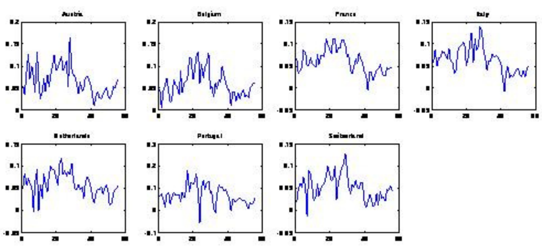
  - 4-cluster structure is appropriate
- The weighting method is then applied
  - ASC of the weighting approach is 0.8024, indicating a clearer structure
  - C1={Austria, Belgium, France, Italy, Netherlands, Portugal, Switzerland}, C2={Cyprus, Denmark, Finland, Sweden, Turkey, United Kingdom}, C3={Iceland, Spain}, and C4={ Ireland, Luxembourg, Norway}
  - Identical to the unweighed clustering solution.

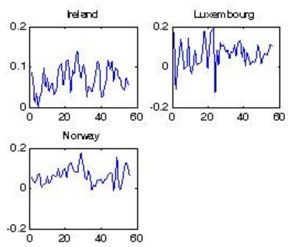
MONASH University
Business and Economics

www.**buseco**.monash.edu

Conclusions

- Consistency of clustering solution with the patterns
  - Example: C1 members and C4 members





- The weighted features show superiority to the unweighted ones
- it is recommended that in applied works, the proposed weighting method should be incorporated with the clustering algorithms

MONASH University
Business and Economics

www.buseco.monash.edu

Thank you!

8. A Methodology for Decomposing Age, Period and Cohort Effects Using seudo-Panel Data to Study Children's Participation in Organised Sports

A Methodology for Decomposing Age, Period and Cohort Effects Using Pseudo-Panel Data to Study Children's Participation in Organised Sports

Australian Statistical Conference

Perth

6 December 2010

Anil Kumar and Peter Rossiter

Presentation Outline

- Aim of study
- Creating Pseudo Panel Data from Repeated Cross-sectional Surveys
- Decomposing Age, Period and Cohort (APC) Effects – Simple Accounting Framework
- Modelling Sports Participation – Logistic Regression
- Conclusions

Aim of study

- Pool repeated cross-sectional surveys to create a pseudo-panel data to study children's sports participation within a longitudinal framework.
- Pseudo-panel data not true longitudinal data so can't provide insights at the individual level, but it can at a group or cohort level.
- Main research questions examined:
  - how sports participation changes over a child's life cycle and factors underlying these changes?
  - is there age, period and cohort effect in sports participation?

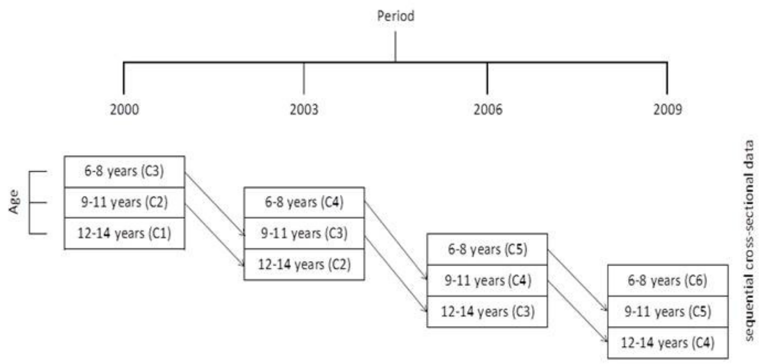
What are APC effects?

- APC analysis useful when there is a time specific or time-related variation in the phenomenon.
- ***Age*** effect relates to changes that occur as group/people age
  - e.g., does rate of participation in organised sports increase, decease or remain unchanged as children grow older over time (5-14 years)?
- ***Period*** effect relates to the influence of the time or period in which the event occurs i.e. variation over time
  - e.g., as a result of growing concern over childhood obesity and public health campaigns/policy initiatives to address this issue is participation in organised sports by children increasing over time?
- ***Cohort*** effect is the effect specific to those born around the same time.
  - e.g., are rates of sports participation of younger cohorts of children higher, lower or the same compared to older cohorts of children?

Data Source

- Survey of Children's Participation in Culture and Leisure Activities (CPCLA)
  - repeated cross-sectional survey covering children population aged 5-14 years conducted every 3 years in April.
  - data on demographics, selected organised sport and cultural activities outside of school hours.
- Focus here on **organised** sports as defined in the survey.
- Four waves pooled here (previously three waves).
  - 2000, 2003,2006, 2009

Construction of Pseudo-panel data Using CPCLA



Cohort	Birth Year
C1	1986-88
C2	1989-91
C3	1992-94
C4	1995-97
C5	1998-00
C6	2001-03

Construction of Pseudo-panel data Using CPCLA

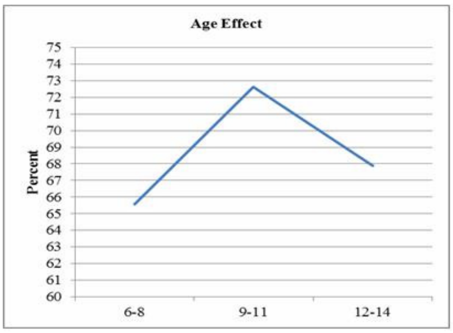
	2000	2003	2006	2009
6-8	C3	C4	C5	C6
9-11	C2	C3	C4	C5
12-14	C1	C2	C3	C4

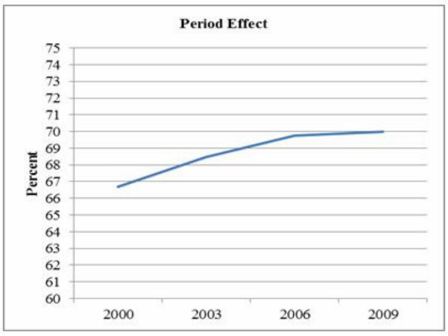
C=A+P-1=3+4-1=6

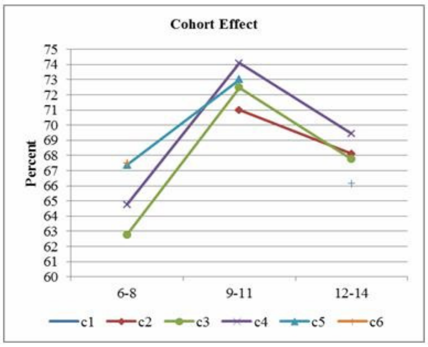
Construction of Pseudo-panel data Using CPCLA

	2000	2003	2006	2009	Total
6-8	62.8	64.8	67.4	67.5	**65.3**
9-11	71.0	72.5	74.1	73.0	**72.6**
12-14	66.1	68.1	67.8	69.4	**67.7**
Total	**66.7**	**68.5**	**69.8**	**70.0**	

Age, Period and Cohort Effects (unadjusted)







APC Effects (unadjusted) (cont)

- The problem with these diagrams is they include the influence of APC effects together which can be confounding.
- Need to decompose these effects separately.
  - How do you do this?

APC Accounting Framework

- Basic APC decomposition accounting model can be written in a linear form as follows:

$$M_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijk}$$

$$M_{ijk} = \ln\left(\frac{R_{ijk}/100}{1-R_{ijk}/100}\right)$$

$$\sum_{i=1}^{3}\alpha_i = \sum_{j=1}^{4}\beta_j = \sum_{k=1}^{6}\gamma_k = 0$$

- M_{ijk} - natural log of the odds of sports participation corresponding to the respective APC cell (12 cells)

•μ – intercept term or the average log-odds or underlying rate pertaining to the complete target population

- The APC effects are parameterised/constrained such that their effects add up to zero.

Identification Problem of APC Modelling

$$M_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijk}$$

- But inclusion of all three variables in the model poses a problem for model estimation.

- APC variables not independent of each other but any one variable is a linear combination of the other two.
  - Given A and P we can determine the birth cohort (C), since
    - C = P – A or P=C+A or A=P-C

- This gives rise to the problem of perfect collinearity or the 'identification ' problem in which it is not possible to simultaneously estimate the true effects of APC, unless some additional constraints are imposed on the parameters of some of these variables.
  - In matrix terminology this yields a singular design matrix of one less than full rank and as such no inverse exists.

Methods for APC Decomposition

- Several methods proposed to resolve this problem
- Such methods generally referred to as coefficients constraints approach
  - Assume only two of the three APC variables affect the outcome
    - generally assume cohort or period effect to be zero on average but keep age in because it is an important determinant of social behaviour.
  - Constrain some parameters to be equal - based on some theoretical argument or observation of data
    - either assume two age, two period, or two cohort parameters are equal.
  - Use proxy variables - assume one of APC represented by some other variable
    - E.g assume cohort effect is proportional to cohort size, or the unemployment rate might be used as a proxy for period effect.

Methods for APC Decomposition (cont)

- All these methods, however, require strong theoretical assumptions and have some issues/problems which may or may not be justified in particular.
  - Element of arbitrariness/value judgement
  - Reliance upon external information
  - Sensitivity of parameter estimates to choice of constraints
- There does not seem to be any consensus as to the most appropriate method to use to resolve this identification problem.

Methods for APC Decomposition (cont)

- If the matrix is of full rank or once constraints are imposed then we can solve for *b*

$Y = Xb + \varepsilon$

➢ (conventional matrix form of least squares regression)

$\hat{b} = (X^T X)^{-1} X^T Y$

b=$(\mu, \alpha_1\ldots, \alpha_{a-1}, \beta_1\ldots, \beta_{p-1}, \gamma_1\ldots, \gamma_{a+p-2})^T$

➢ model parameters

Intrinsic Estimator Approach

- A more satisfactory solution to this APC identification problem has been proposed by Yang *et al in their* 2008 paper

"The Intrinsic Estimator for Age-Period-Cohort Analysis: What It Is and How to Use It"

- Method referred to as the Intrinsic Estimator (IE).
- Their solution is based upon the Moore-Penrose generalised inverse.
  - ➢ method for finding inverse when matrix is singular or of not full rank
- Given matrix *X* its generalised inverse X^+ produces a unique solution to the least squares equation:
  $\hat{b} = X^+ Y$
- Derivation of the IE is equivalent to conducting a principal components regression analysis, and applying an inverse transformation to the parameter estimates to recover the age, period cohort interpretation.

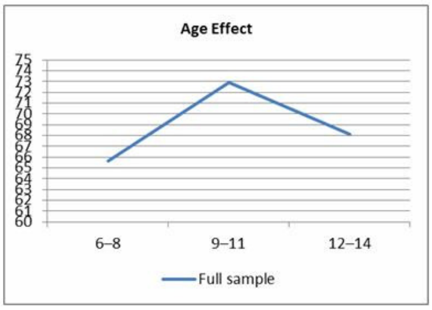
IE Approach (cont)

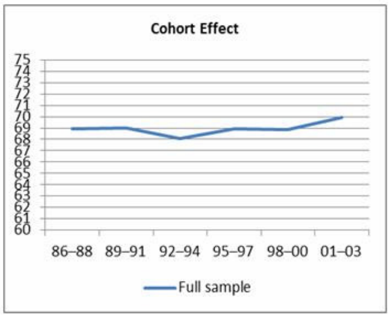
- In SAS can use ***proc iml*** to compute the generalised inverse and derive the $\hat{b}$ coefficients quite easily.
- Method removes the arbitrariness in the choice of coefficient constraints i.e. it restores objectivity to the analysis in that it lets the data decide the shape of the effects.
- Method can be shown to have certain desirable properties (lower bias/variance) with respect to the coefficient constraints solutions.

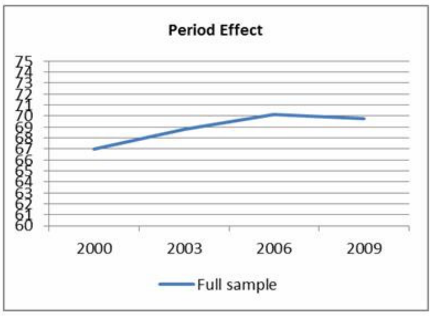
Results based on IE Approach

- We can use the IE method to decompose the APC effects separately as shown in the following charts.

IE – APC – Full Sample







Logistic Modelling

- We can now use information from the simple accounting framework based on the IE approach, as well as add other variables, in addition to APC, to estimate a ***fuller*** model for sports participation.

- Here we are interested in finding out what variables influence children's participation in sports in addition to APC effects.

- We can use a logistic modelling framework to examine this.

Model variables

- Explanatory variables cover demographic, geographic and socioeconomic variables as well as dummy variables for age, period and cohort .

- Non-APC variables
  - sex
  - family status
  - migrant status
  - geographic location
  - parents' employment status
  - SEIFA (3)
  - above/below average TV/computer use

Model variables (cont)

- APC variables

  - Use information from earlier IE analysis as a guide to how to enter the APC variables in the model.

  - Since cohort effect was found to be 'insignificant' we can combine or collapse some of the categories for this and focus on estimating the stronger age and period effects.

    - Age (3) – same three age groups
    - Period (4) – same four periods
    - Cohort (4) – collapse from 6 to 4 categories

  – C1&C2, C3, C4, C5&C6

Model Results

Parameter	Estimate	SE	P Value	Odds Ratio
Intercept	0.923	0.0491	<.0001	
Aged 9-11 years	0.3736	0.0541	<.0001	1.453
Aged 12-14 years	0.1283	0.0876	0.1428	1.137
2003 Survey	0.1053	0.0544	0.0528	1.111
2006 Survey	0.2036	0.0963	0.0344	1.226
2009 Survey	0.2203	0.1277	0.0846	1.246
Cohorts 1&2 - Born 1986-1991	0.0724	0.0672	0.2812	1.075
Cohort 4 - Born 1995-1997	0.0149	0.0565	0.7926	1.015
Cohort 5&6 - Born 1998-2003	-0.00362	0.1064	0.9729	0.996
Girls	-0.2547	0.0282	<.0001	0.775
Both parents born overseas	-0.6636	0.0404	<.0001	0.515
Living in rest of the state	0.0446	0.0346	0.1977	1.046
Single parent family	-0.1444	0.0407	0.0004	0.866
No parent(s) in employment	-0.8183	0.0452	<.0001	0.441
Highest SEIFA quintile	0.5829	0.0457	<.0001	1.791
Lowest SEIFA quintile	-0.5117	0.0429	<.0001	0.599
Above average television and computer usage	-0.1634	0.0393	<.0001	0.849
N	29814			
Likelihood Ratio p value	<.0001			
Hosmer-Lemeshow goodness-of-fit p value	0.6098			
Max-rescaled R-Square	0.1161			
% Concordant	67.1			

Model results (cont)

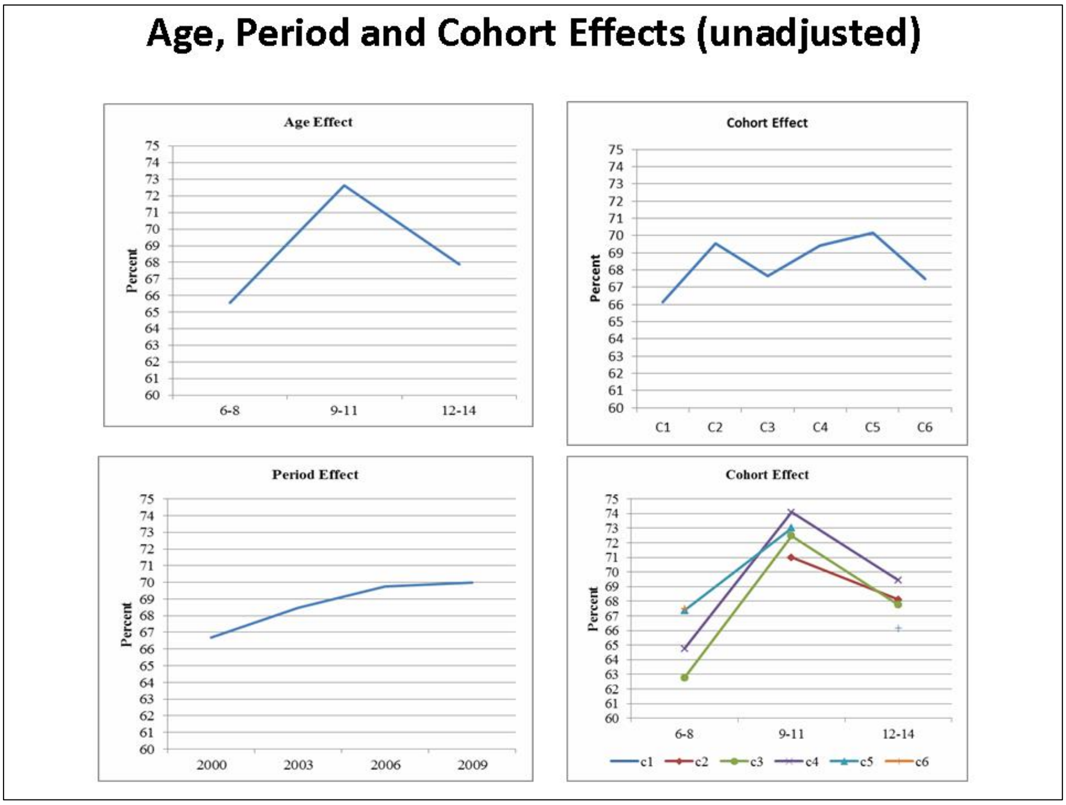
- Model results are expected for the non-APC variables
- 7 of the 8 variables here are stat significant at 1% level.
  - sex, parents' cob & employment status, socioeconomic status, time spent on TV/computers and family status appear to be strongly associated with sports participation by children.
  - Area (city vs rest of state not significant)
- Age effects – those 9-12 significantly different compared to 6-8 but those 12-14 not significantly different from those 6-8.
- Period effects – monotonic trend i.e rising over time (significant @3-8% level)
- Cohort effects not found to be significant, confirming earlier analysis.
- Four waves results similar to 3 waves results but one age group (12-14) and area variable no longer significant.

Conclusions

- Estimates obtained by the IE method are more direct and do not require prior information to select appropriate model identifying constraints.
- Its practical value is to provide objective evidence of the relative effects of age, period and cohort in standard application.
- Method may well provide a useful alternative to conventional methods for the APC analysis.
  - Handy in cases where there are more period and time categories and little information/guidance on where to impose the constraints.

References

- Yang, Y.; Schulhofer-Wohl, S.; Fu, W.J. and Land, K.C. (2008) "The Intrinsic Estimator for Age-Period-Cohort Analysis: What It Is and How to Use It", *American Journal of Sociology*, 113(6), pp. 1697–1736.
- Kumar, A.; Rossiter, P. and Olczyk, A. (2009) Children's Participation in Organised Sporting Activity, Research Paper, 1351.0.55.028, Australian Bureau of Statistics, Canberra.


Age, Period and Cohort Effects (unadjusted)
Age Effect
Cohort Effect
Period Effect
Cohort Effect
Percent
6-8
9-11
12-14
C1
C2
C3
C4
C5
C6
2000
2003
2006
2009
c1
c2
c3
c4
c5
c6

9. School segregation, class size and student achievement patterns in Salvador de Bahia (Brazil)

School segregation, class size and student achievement patterns in Salvador de Bahia (Brazil)

Paulo A. Meyer M. Nascimento
Research officer at the Brazilian National Institute for Applied Economic Research - IPEA
paulo.nascimento@ipea.gov.br

AUSTRALIAN STATISTICAL CONFERENCE 2010
Fremantle, Australia, 6th December 2010

School segregation, class size and student achievement patterns in Salvador de Bahia (Brazil)

The focus of the study

- School segregation patterns;
- Student achievement patterns;
- The association between class size and student achievement.

School segregation, class size and student achievement patterns in Salvador de Bahia (Brazil)

The dataset

- 55 schools, 169 classes and 4,025 students;
- That is a representative sample of urban schools located in Salvador de Bahia – the third biggest Brazilian city in population;
- Students were enrolled at their first year of primary school.
- Data refers to the first of a 4 year project that took place in 5 big Brazilian urban centres.

School segregation, class size and student achievement patterns in Salvador de Bahia (Brazil)

Instruments of data collection

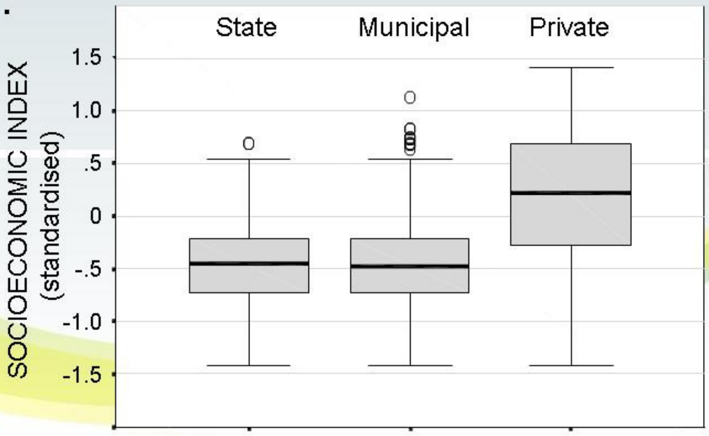
- A Reading baseline test;
- A questionnaire applied to the teachers;
- A questionnaire applied to the families;
- A Reading test at the end of the academic year.

School segregation, class size and student achievement patterns in Salvador de Bahia (Brazil)

ipea

School segregation patterns (I)

- In Salvador, public schools are solely to the very poor:

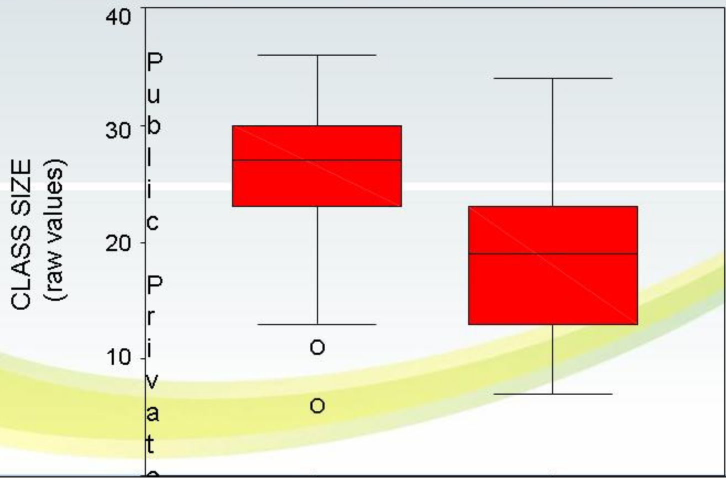


School segregation, class size and student achievement patterns in Salvador de Bahia (Brazil)

ipea

School segregation patterns (II)

- Classes are more crowded at public schools:



School segregation, class size and student achievement patterns in Salvador de Bahia (Brazil)

Student achievement patterns

- When only the Reading test undertaken at the end of the academic year is accounted for, T-test for equality of means shows a .86 standard-deviation-point difference between the scores of private and public school students;
- The difference disappears when progress between the two tests is the variable of interest.

School segregation, class size and student achievement patterns in Salvador de Bahia (Brazil)

Class size and student achievement (i)

- A multimodel value-added model was applied:

$$A_{pcs1} = \alpha + \theta A_{pcs0} + \beta Z_{cs} + \phi T_{cs} + \psi F_{pcs} + \varphi I + \nu_s + \upsilon_{cs} + \in_{pcs}$$

- *Apcs1* is *Achievement*, i.e. the test score at wave 2 of pupil *p* in class *c* in school *s*;
- *Apcs0* is *Initial ability*, i.e. the test score at wave 1 of pupil *p* in class *c* in school *s*;
- *Zcs* is the natural log of the size of class *c* in school *s;*
- *Tcs* is a vector comprising teacher characteristics;
- *Fpcs* is a vector for family inputs;
- *I* is the vector of interaction terms;
- νs, υcs and ϵpcs are residuals for schools, classes and pupils, respectively.

School segregation, class size and student achievement patterns in Salvador de Bahia (Brazil)

Table Estimations for the coefficients and respective standard errors

Variable	Estimate	Standard error
Fixed		
(constant)	3.496	0.297
Initial ability	0.514	0.014
log class size	0.014	0.093
Parental assets	-0.006	0.016
Parental education	0.082	0.015
Teacher schooling	0.026	0.062
Teacher experience	0.008	0.028
School type	0.422	0.107
Random		
Pupil-level variance	0.581	0.014
Class-level variance	0.062	0.012
School-level variance	0.034	0.014
-2* loglikelihood	8715.820	

School segregation, class size and student achievement patterns in Salvador de Bahia (Brazil)

<u>Class size and student achievement (ii)</u>

- Results show no associations between class sizes and student achievement;
- Causal effect links could not be tracked, because of the lack of defensable instrumental variables.

School segregation, class size and student achievement patterns in Salvador de Bahia (Brazil)

Discussion on instrumental variables for estimating class size effects on student achievement

- The use of the IV approach to tackle the endogeneity problem of class sizes depends on the availability of longitudinal data or the existence of an institutional rule or arrangement that generates discontinuities in the sample.
- Instruments such as school average class size or school size are dismissed.

School segregation, class size and student achievement patterns in Salvador de Bahia (Brazil)

Future developments

- With the full data set on hand, try to tackle the endogeneity problem using an IV based on natural variations (Hoxby, 2000).
- Alternatively, a PSM could be plausible with the complete data set.


School segregation, class size and student achievement patterns in Salvador de Bahia (Brazil)
ipea
THANK YOU!
Paulo A. Meyer M. Nascimento
Research officer at the Brazilian National Institute for Applied Economic Research - IPEA
paulo.nascimento@ipea.gov.br

10. Merging Colorectal Cancer Surveillance Data, an experience report

"Relax Mrs Flanagan, I've never know a high fibre exorcism fail yet!"

www.csiro.au

Merging Colorectal Cancer Surveillance Data, an experience report.

ASC 2010: Statistics in the West, Fremantle, WA

Norm Good
John O'Dwyer, Russell Diehl, Finlay Macrae, Graeme Young, Bill Venables, Masha Slattery

7 Dec 2010

National Research FLAGSHIPS Preventative Health
CSIRO

Merging Surveillance Datasets(MSD)

Aims

1. Identify factors predicting findings at colonoscopy
2. Quantify sensitivity and specificity of Faecal Occult Blood Tests (FOBT) in symptomatic and asymptomatic persons
3. Identify ideal interval between colonoscopies for identified risk groups

MSD- a tale of two cities

National Research FLAGSHIPS Preventative Health CSIRO

Data background

- Integrate data from two large surveillance programs to provide a larger data set to address the project aims
  - Development of common standard for fields required in analysis
  - Agreed by 'opposing' clinicians
  - To cater for differing philosophies of data capture
  - To meet working definitions of standards
  - To address the challenges of data entered to run a surveillance program vs for research
- Data derived from The Royal Melbourne Hospital (BCSP) and Southern Adelaide Health Services (SAHS) Bowel Cancer Screening Program
- The programs have been prospectively planning and documenting screening in familial bowel cancer for 29 and 16 years respectively

MSD- a tale of two cities

National Research FLAGSHIPS Preventative Health CSIRO

What's so difficult about merging data?

- Two large data sources from 3 hospitals
  - Royal Melbourne Hospital (Prof Finlay Macrae)
  - Flinders Medical Centre, Repatriation General Hospital (Prof Graeme Young)
- Both record surveillance data
- Both measure outcomes

"So pulling this data together should be easy right?"

National Research FLAGSHIPS Preventative Health CSIRO

Follow the rules and warm fuzzy motherhood statements

- Kelman, Bass & Holman, *Research use of linked health data – a best practice protocol*, Aust NZ J Public Health 2002; 26:251-5

"A protocol for **facilitating access** to administrative data from multiple organisations for the purpose of health services research"

"**Promotes confidence** within the community & data custodians that linked health information is simultaneously delivering research benefits and rigorously protecting individual privacy"

"In Australia, an additional barrier to research is the result of health data sets being collected by different levels of government – thus all are not available to any one authority"

"Health data, although widely and diligently collected, continue to be under-utilised for research and evaluation in most countries"

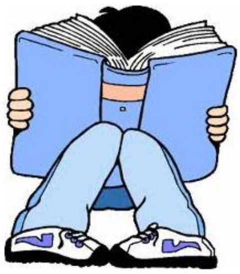
National Research FLAGSHIPS Preventative Health CSIRO

The Usual Data Cleaning Extravaganza

- Iterative process of identifying field level and related-field level data errors
- Everything from typos to inconsistent data entry to surveillance program issues, eg:
  - hitting 1 instead of 2 (and thereby recording a result of adenoma instead of cancer)
  - Entering all incoming patients with a familial risk of 'CAS' – even if they don't have cancer
- Checks performed:
  - Valid dates of birth
  - Codified fields contain only valid entries
    - Gender (0,1) Adelaide, (1,2) Melbourne
  - Colonoscopy results matches pathology result
  - Individual status summary matches colonoscopy result
  - Blank or duplicate entries
  - Young ages at colonoscopy correct
- Every entry in every field validated – in value and context
- Enormous effort by surveillance program staff to go back to paper to check results, and to enter new / changed data

National Research FLAGSHIPS Preventative Health CSIRO

BUT, and this is where the story really starts



MSD- a tale of two cities

Example 1: Polyp Size

"Simple transformation"

Melbourne has values L (large) and S (small)

Adelaide translates

>=10mm = L

<10mm = S

This requires both clinicians to agree on the standard once

Example 2: Advanced Adenoma

"Complex Transformation"

Captured in Adelaide as "Advanced Adenoma"

Requires transformation in Melbourne:

Advanced Adenoma = number of polyps >=3 or
any large adenoma or
high grade dysplasia or
significant villous change or
serrated adenoma

Simple once you know how, but getting a concise definition requires iterative clinical and data manager input from both sites

National Research FLAGSHIPS Preventative Health CSIRO

Example 3: Family History & Personal History

"Philosophical Difference"

Captured at both sites in a similar way

BUT!

Prof Graeme Young says:

"As we learn more about the patient we uncover their genetic predisposition to cancer, so their family history becomes more and more accurate"

i.e. We only store the most recent family history

Prof Finlay Macrae says:

"The family history over time tells us the story of why we choose a particular patient treatment plan"

i.e. We only store family history for each surveillance event

Both valid view points but very different ways of capturing data

National Research FLAGSHIPS Preventative Health CSIRO

Whilst all that stuff went on.......

Small dataset from Adelaide, N~1900

- Data was audited in greater detail
- Aim to look at effectiveness of interval FOBT testing

Major results

- Risk of presenting at colonoscopy with a significant neoplasia reduced by 65% if you undergo interval FOBT testing.
- Median reduction in delay to diagnosis for cancer was 26 months, and for advanced adenoma was 18 months
- Positive FOBT detected 12/14 (86% sensitivity) cancers and 60/96 advanced adenomas (63%)

Lane, J.M., Chow, E., Young, G.Y., Good, N.M., Smith.A., Bull, J., Sandford, J., Morcom, J., Bampton, P.A. and Cole, S.R. (2010). Interval fecal immunochemical testing in a colonoscopic surveillance program for increased risk for colorectal cancer, *accepted Gastroenterology*.

National Research FLAGSHIPS Preventative Health CSIRO

Given all the differences so far how does the data look?

- Are inferences we make going to be affected by the data source?

- Do some exploratory data analysis on what we think are important factors

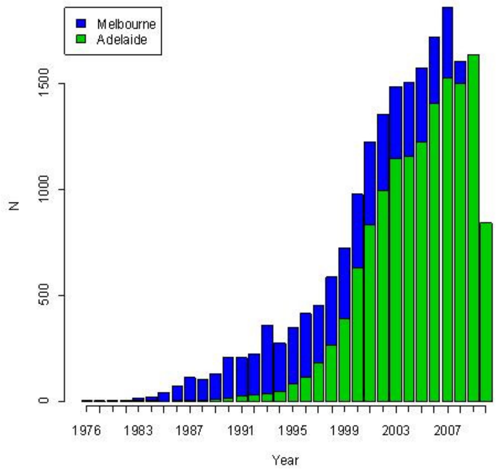
National Research FLAGSHIPS Preventative Health CSIRO

Number of records	60151
Number of patients	10499
Age	58.9(13.2)
Male	43.50%

MSD- a tale of two cities

National Research FLAGSHIPS Preventative Health CSIRO

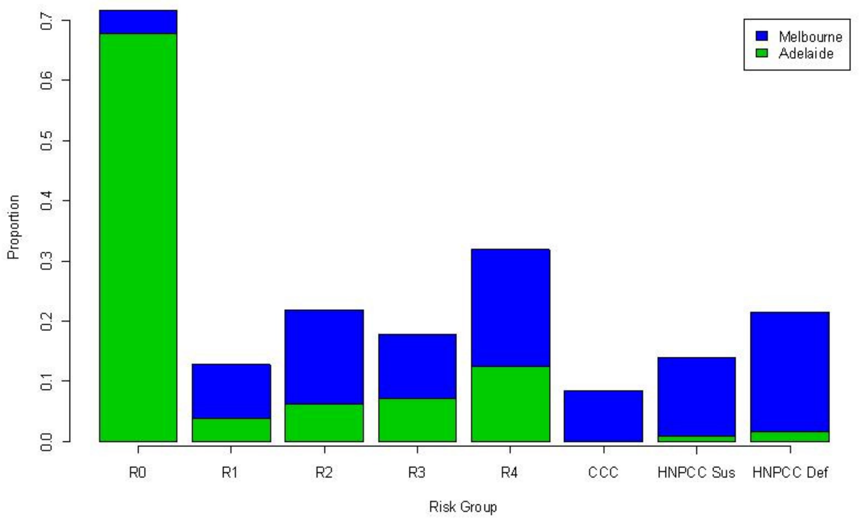
Look at important variables and interactions with data source



MSD- a tale of two cities

National Research FLAGSHIPS Preventative Health CSIRO

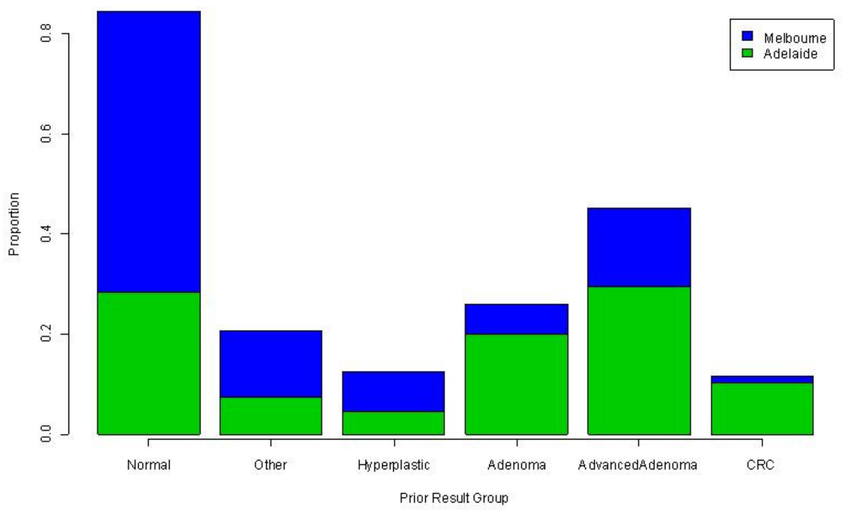
Family history of CRC



MSD- a tale of two cities

National Research FLAGSHIPS Preventative Health CSIRO

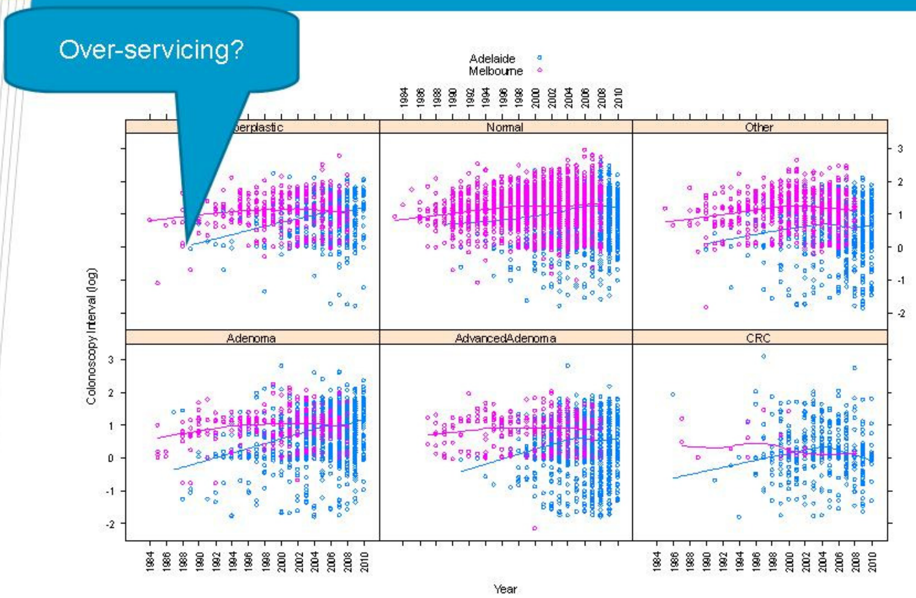
Previous colonoscopy findings



MSD- a tale of two cities

National Research FLAGSHIPS Preventative Health CSIRO

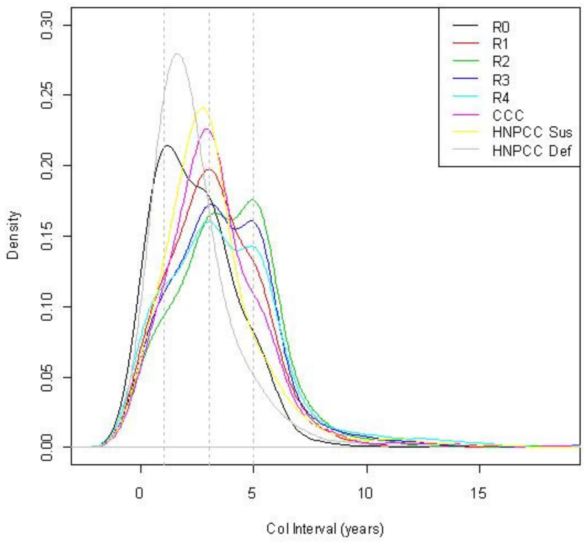
And you can even have a bit of fun with clinicians



MSD- a tale of two cities

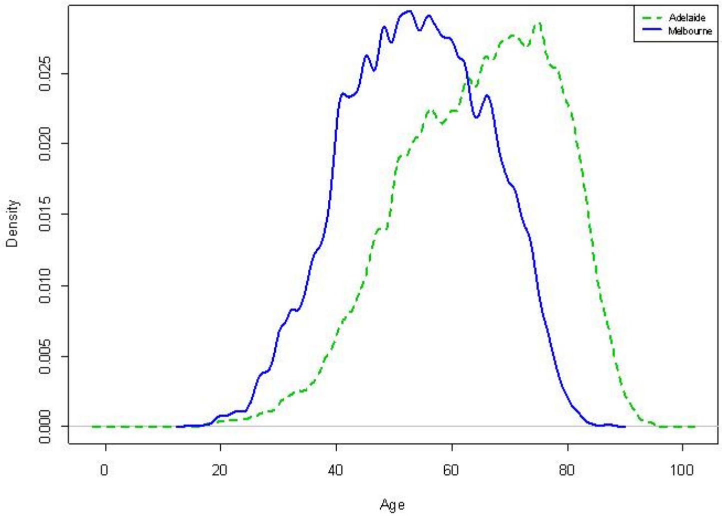
National Research FLAGSHIPS Preventative Health CSIRO

NO, R0 group just followed up more rigorously



MSD- a tale of two cities

National Research FLAGSHIPS Preventative Health CSIRO



MSD- a tale of two cities

National Research FLAGSHIPS Preventative Health CSIRO

What does this mean in practise?

1. Explore interactions (none significant so far)

	Df	Deviance	Resid.DF	-2*LL	P(>|Chi|)
Age	1	381.1037	5706	6678.981	0.00
Sex	1	15.491	5705	6663.49	0.00
RiskRelAlt	7	38.64511	5698	6624.845	0.00
priorResult	5	274.5312	5693	6350.314	0.00
DataSource	1	0.586741	5692	6349.727	0.44
priorResult:DataSource	5	5.454845	5687	6344.273	**0.36**

2. Look for potential biases in the data

ASSUME NOTHING

MSD- a tale of two cities

National Research FLAGSHIPS Preventative Health CSIRO

Take home message....as always ☺

- Relationship management
- Clinical data management
- Clinical data analysis
- Data cleaning
- Statistical analysis
- Project management

MSD- a tale of two cities

National Research FLAGSHIPS Preventative Health CSIRO

Acknowledgements

Clinical

- Prof. Finlay Macrae (Royal Melbourne Hospital)
- Prof. Graeme Young (Flinders Medical Centre)

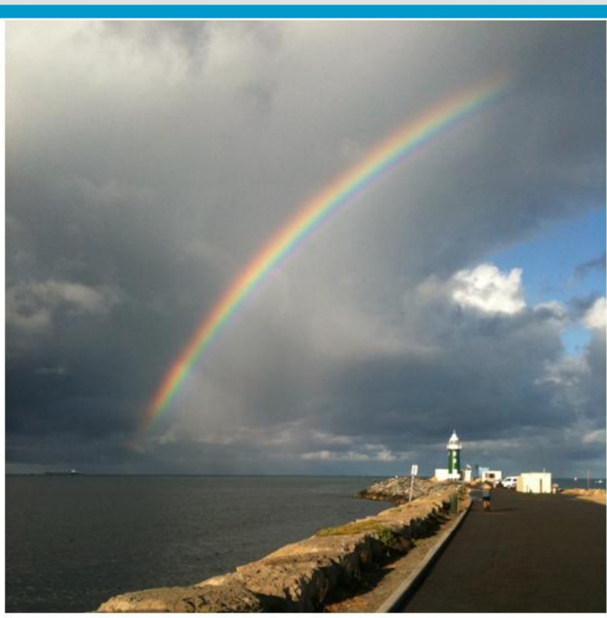
Data collection and management

- Masha Slattery (Royal Melbourne Hospital)
- Joanne Lane (Bowel Health Service, Adelaide)

MSD- a tale of two cities

National Research FLAGSHIPS Preventative Health CSIRO

Thank you



MSD- a tale of two ciities

National Research FLAGSHIPS Preventative Health CSIRO

11. Multivariate statistics in tax administration

Inland Revenue
Te Tari Taake

Multivariate statistics in tax administration

Bhaskaran Nair, Graeme Buckley, Michael Slyuzberg and Xin Wang

7 December, 2010

National Research Unit
New Zealand Inland Revenue
Wellington

Overview

- National Research Unit- Who we are, what we do?
- Application of multivariate and data mining techniques using tax data
- Issues
- Questions

NaRU - Who We Are, What we do?

- Happy bunch of 15 researchers: statisticians, sociologists, psychologists and data analysts.
- Part of Corporate Strategy Group within NZ Inland revenue
- Centre of excellence for research and statistical analysis

work plan focusing on:

- Understanding compliance behaviour
- Profiling different customer groups to better target information or interventions, and
- Ensuring the robust collection of customer satisfaction and community perceptions data

Multivariate Analysis using Tax Data

➢ We cover three projects outlining their objectives, methods, data sources, issues and challenges

- **Understanding compliance behaviour**
  - ***Identification of key drivers***
  - ***Compliance Dynamics***
- **Segmentation of customers**
- **Influence of external factors on tax compliance**
  - ***Economic factors***
  - ***Social factors***

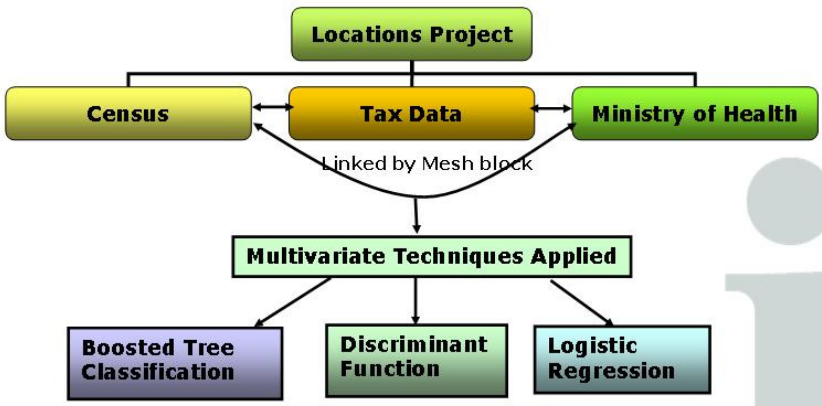
Understanding Compliance Behaviour -1

Project Objectives:

- Investigate differences in customers' filing and payment behaviour by geographical location
- Analyse the rationale behind those differences;
- focussing on targeted customer groups based on their filing and payment behaviour

Inland Revenue Te Tari Taake ASC 2010 Perth 단기 4344년 1월 3일 5

Understanding Compliance Behaviour -2 Methodology Design

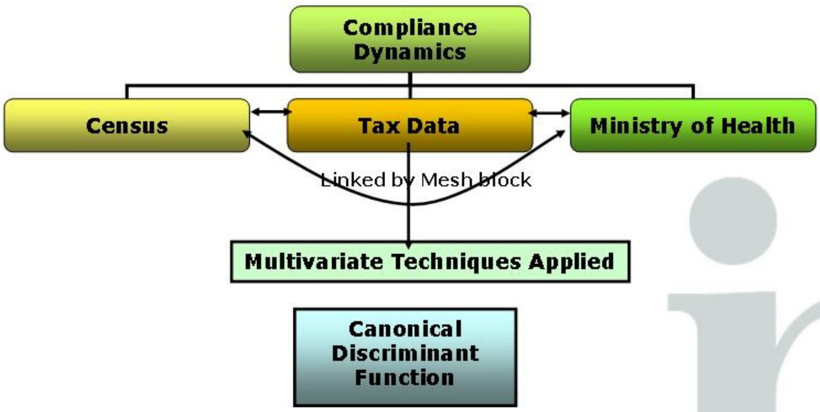


Inland Revenue Te Tari Taake ASC 2010 Perth 단기 4344년 1월 3일 6

Understanding Compliance Behaviour –3
Classification summary payment compliance

Multivariate techniques	**Correct classification 2001(%)**		**Correct classification 2006(%)**		**Overall Error(%)**	
Models	**Non Compliance**	**Full Compliance**	**Non Compliance**	**Full Compliance**	**2001**	**2006**
Boosted classification tree	68	53	68	54	40	39
Logistic Reg	1	100	7	100	50	47
Disc Fn kNN	27	98	31	98	38	36

Understanding Compliance Behaviour –4
Compliance Dynamics -Methodology Design

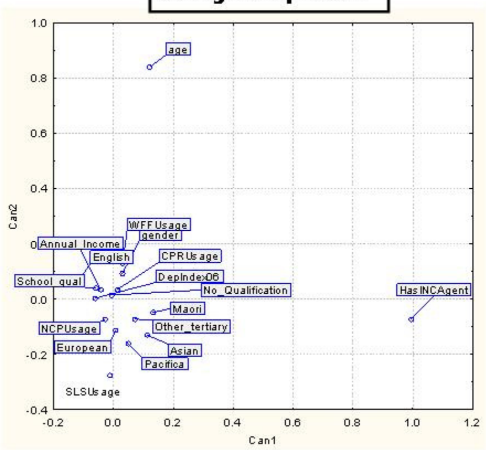


- **Compare Compliance behaviour over time**
- **Identify drivers of change - Three types of changes**

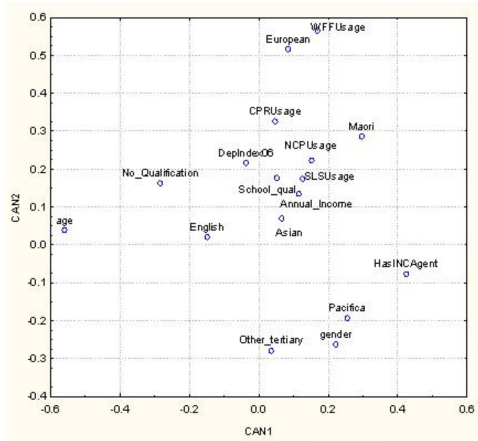
Understanding Compliance Behaviour –5

Canonical scatter plot on Filing and Payment Compliance Dynamics

Filing Compliance



Payment Compliance

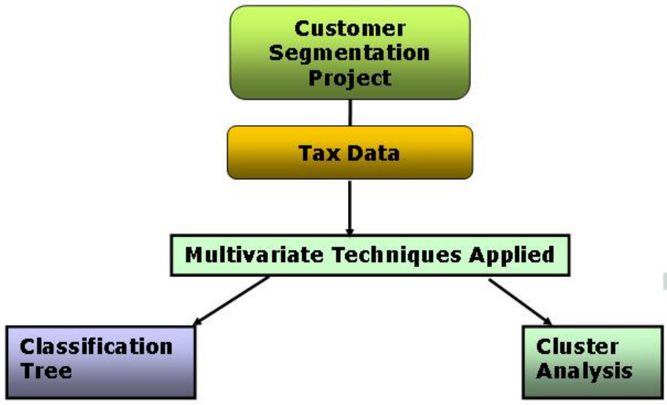


Segmentation of Customers -1

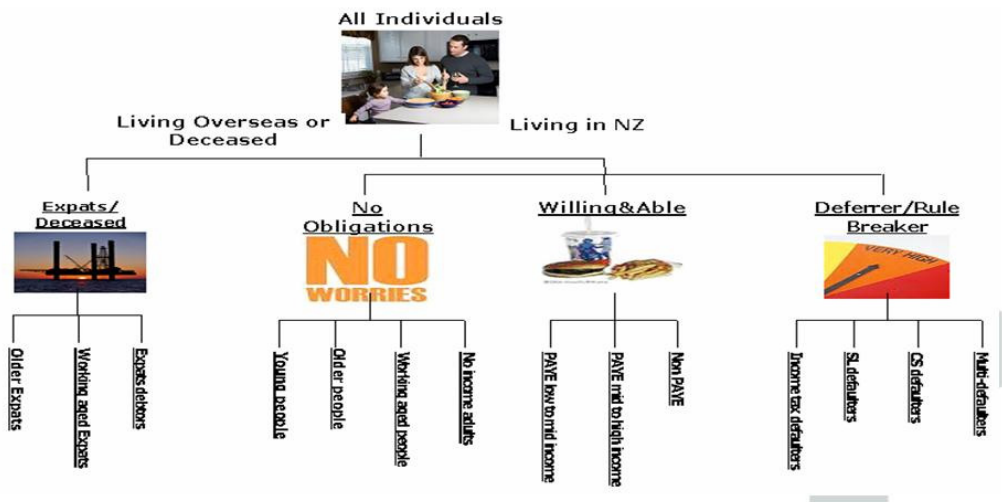
- understand customers and their behaviour

- Data diversity;
- Feature selection;
- Business expectation;
- Clustering architecture
- Segmentation technique;

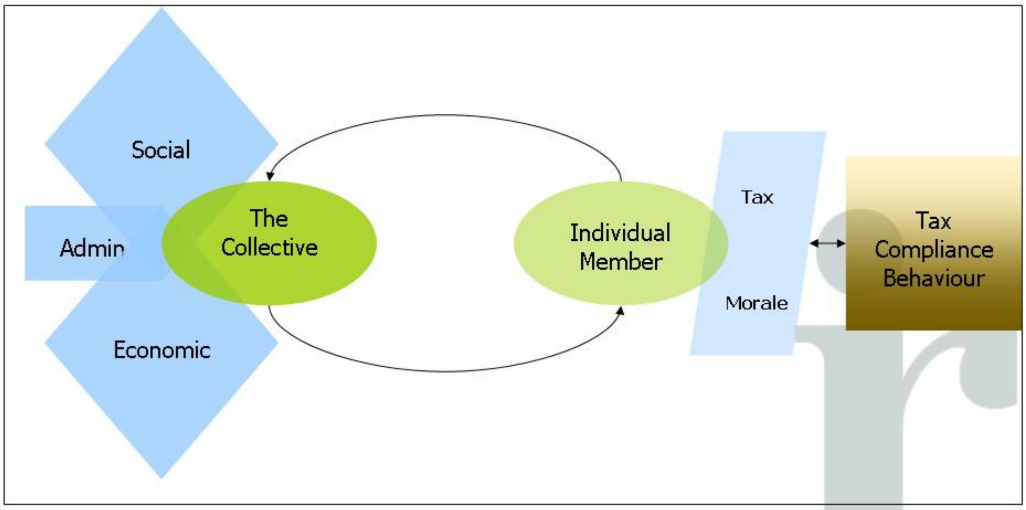
Segmentation of Customer -2
Methodology Design



Segmentation of Customers -3
Understanding customers and their behaviour

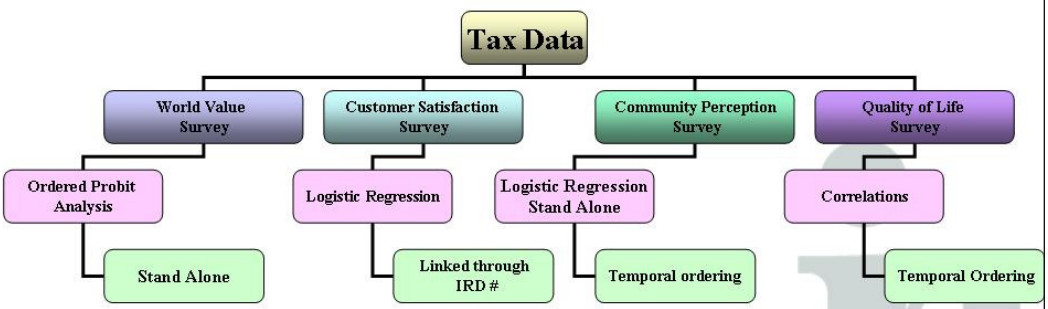


External factors -1
Model for Influence on Tax Compliance



Inland Revenue
Te Tari Taake
ASC 2010 Perth
단기 4344년 1월 3일
13

External Factors -2
Methodology Design

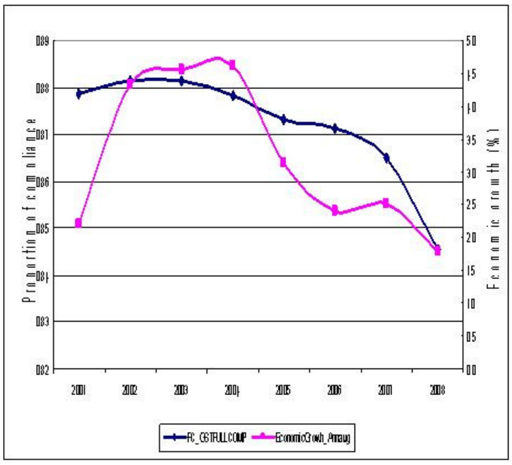


Inland Revenue
Te Tari Taake
ASC 2010 Perth
단기 4344년 1월 3일
14

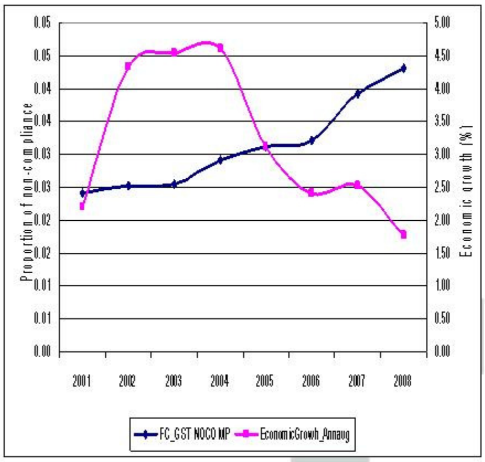
External Factors -3

Effect of Economic Factor on Compliance Behaviour

On-time Filing Compliance



Non-Compliance



Issues -1

specific to tax data

- Supervised learning tasks
- Class Imbalance
- Missing data is generally not missing at random
- The level of sophistication in documented business processes
- Frequent changes to policy or business process
- Some fields in tax records are sparsely populated

Issues -2
General

- ➢ Confounding factors
- ➢ Assumption of regression analysis
- ➢ Multicolinearity
- ➢ Specification error
- ➢ Measurement error
- ➢ Nominal and ordinal data can not be used in Factor analysis

Issues -3
How we tackled the challenges?

Challenges

- Supervised Learning task
- Class Imbalance
- Sparse/Missing
- Administrative Data
- Survey Data
- Data Integration
- Data Diversity
- Qualitative and Quantitative data

Resolutions

- Oversampling minority class
- Multiple imputation for missing data
- Link data using common factors
- Selection of appropriate analytic techniques
- Mixed Methods

Questions?

Inland Revenue
Te Tari Taake

ASC 2010 Perth 단기 4344년 1월 3일 19

12. Sampling in the real world

Sampling in the real world

By
Martin Caruso
Australian Statistical Conference
8th December 2010

Overview of talk

- Background on the AWE survey
- The impetus for change
- AWE parallel sample run
- Decomposing Movement/Change
- Changing constantly to stay relevant

Background

- Average Weekly Earnings, plays a role in the indexation of pensions, one of many data sources RBA uses, also topical in regards to the differences in male/female wages
- Total employment and earnings are collected from businesses in a typical week and the rate for a particular level obtained by dividing total earnings ($\hat{Y}$) by total employment ($\hat{X}$)

Background continued

- Users tend to focus on the quarterly to quarterly movements
- Businesses selected using synchronised random sampling with rotation to minimise respondent burden
- AWE uses ratio weights

Updating Industry

- Australian New Zealand Industry Classification (ANZSIC) 1993 was created about a year after the internet going mainstream
- The new ANZSIC06 industry classification came out in 2006 and has Information and Communication as a separate category

Industry Division

- In stratification we use industry division for example mining, manufacturing etc
- ANZSIC93 division has 17 divisions while ANZSIC06 has 19 divisions
- As state, sector and size included in the AWE stratification an extra 2 divisions leads to many strata

The big bang approach

- In addition to the dire need to update industry, there was also a need to fix under-coverage issues on the frame, update the auxiliary size variable and a sample redesign
- Rather than cause potentially several disruptions to the time series the ABS decided to do it all at once

Sample design

- Sample redesign was required to keep new sample design efficient, due to the extra 2 A06 divisions
- The old frame was dual coded with ANZSIC93 and ANZSIC06 industry
- The 3 year delay in implementing ANZSIC06 was due to needing a few years of dual coded data for the sample design

AWE parallel sample

- 2 samples with overlap maximised were run in parallel
- Old sample on old frame, old A93 industry classification
- New sample used new frame and A06 industry classification

Aim of parallel sample

- Major aim of the parallel run was to provide a measure of the shift in level estimate
- Used by time series to create a historical A06 series by back-casting
- Conducted in May 2009 and used in checking the A06 movements in the August 2009 publication

The results

- At the state level for 2 states there were big differences
- Given the biggest change to the estimates should really only have come from the industry level
- Following table shows the estimates we published

Table showing the 2 states

State	March 2009 A93 estimates	May 2009 A93 estimates	May 2009 A06 estimates	August 2009 A06 estimates
	Average Weekly Ordinary Time Earnings Full-time Adult males			
WA	$1,470.70	$1,488.80	$1,401.30	$1,424.30
	Average Weekly Earnings Full-time Adult females			
Tasmania	$1,011.90	$1,028.40	$944.70	$974.90

Decomposing Movement/ Change

Australian Bureau of Statistics

$$AWE_{2c} - AWE_{1c}$$

$$= \frac{\sum_{i \in c} w_{2i} y_{2i}}{\sum_{i \in c} w_{2i} x_{2i}} - \frac{\sum_{i \in c} w_{1i} y_{1i}}{\sum_{i \in c} w_{1i} x_{1i}}$$

where c is the level of interest eg Australia

y_{2i} is the gross earnings for unit i at time t2

x_{2i} is the employment for unit i at time t2

w_{2i} is the ratio weight for unit i at time t2

y_{1i} is the gross earnings for unit i at time t1

x_{1i} is the employment for unit i at time t1

w_{1i} is the ratio weight for unit i at time t1

Decomposing Movement/ Change

Australian Bureau of Statistics

$$\sum_{i \in c} \frac{w_{2i}\left(y_{2i} - y_{1i}\right) - A\hat{W}E_{1c}\left\{w_{2i}\left(x_{2i} - x_{1i}\right)\right\} + y_{1i}\left(w_{2i} - w_{1i}\right) - A\hat{W}E_{1c}\left\{x_{1i}\left(w_{2i} - w_{1i}\right)\right\}}{\hat{X}_{2c}}$$

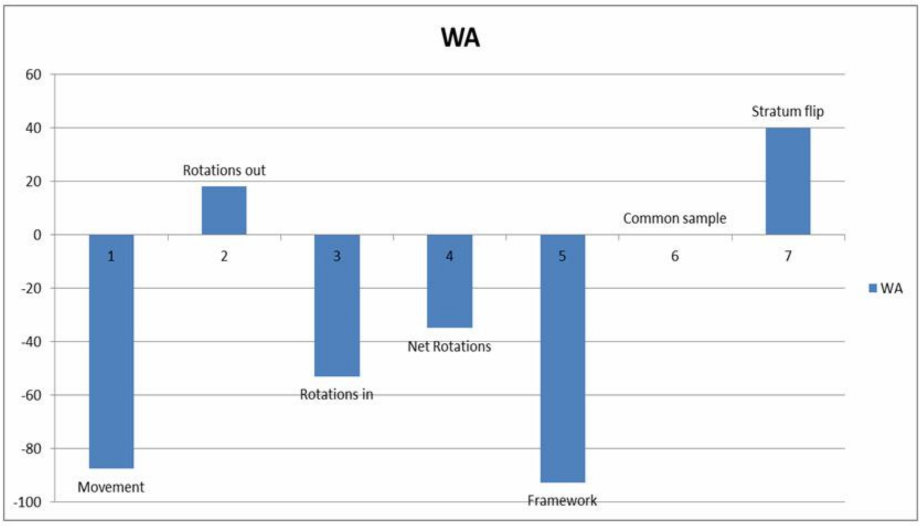
Breaking it down

Main Components	Link to Equation
Common Unit/Stratum Flip (comm)	$\sum_{i\in c\cap comm} \frac{w_{2i}\left(y_{2i}-y_{1i}\right)-A\hat{W}E_{1c}\left\{w_{2i}\left(x_{2i}-x_{1i}\right)\right\}+y_{1i}\left(w_{2i}-w_{1i}\right)-A\hat{W}E_{1c}\left\{x_{1i}\left(w_{2i}-w_{1i}\right)\right\}}{\hat{X}_{2c}}$
Framework (fw)	$\sum_{i\in c\cap fw} \frac{y_{1i}\left(w_{2i}-w_{1i}\right)-A\hat{W}E_{1c}\left\{x_{1i}\left(w_{2i}-w_{1i}\right)\right\}}{\hat{X}_{2c}}$
Rotation in/Birth (rib)	$\sum_{i\in c\cap rib} \frac{w_{2i}y_{2i}-A\hat{W}E_{1c}w_{2i}x_{2i}}{\hat{X}_{2c}}$
Rotation out/Death (rod)	$\sum_{i\in c\cap rod} \frac{A\hat{W}E_{1c}w_{1i}x_{1i}-w_{1i}y_{1i}}{\hat{X}_{2c}}$

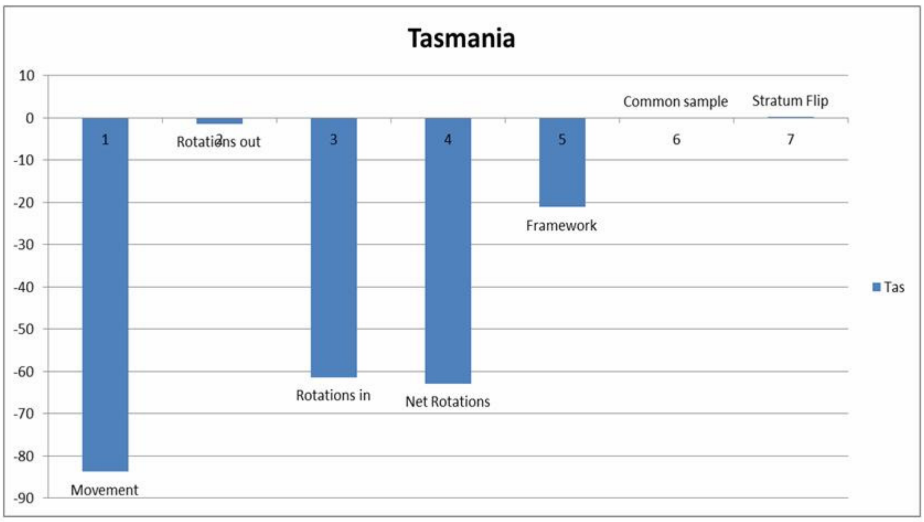
Unit level listings

- Using the equations in the previous table can produce a unit listing in descending order of absolute effect
- Units at the top that stand out from units below can be considered for outliering

The breakdown for WA



The breakdown for Tasmania



Changing to stay relevant

Australian Bureau of Statistics

- ABS will be updating auxiliary variables used in Generalised regression/ratio estimation more frequently
- Long term aim to update industry classification on a more frequent basis
- This implies sample re-designs to keep our sample designs efficient

13. Trustworthy statistics - A shared responsibility?

Trustworthy statistics – a shared responsibility?

Professor Denise Lievesley

Head of School of Social Science and Public Policy,

King's College London *and*

Chair, European Statistical Advisory Committee

1

Themes

- Importance of official statistics for evidence based policy
- Threats to the (perceived) integrity of official statistics
- Statisticians' responsibilities as a professional community working together to build trust

Importance of official statistics

- good government and the delivery of public services
- decision making in all sectors of society
- empowerment of the general public
- democracy
  - providing Parliament and the public with a window on society and the economy, and on the work and performance of government

Statistics fundamental for evidence-based policy

- Helping people to make well-informed decisions about policies, programmes and projects, by putting the best available evidence from research at the heart of policy development and implementation
- Enlightening through making explicit what is known through scientific evidence and importantly what is not known

In contrast to opinion-based policy …

- which relies heavily on
  - either the selective use of information or
  - on the untested views of individuals or groups often inspired by ideological standpoints, prejudices or speculative conjecture
- and policy-based evidence!

Need an evidence base at all stages in the policy cycle

- in shaping agendas
- in defining issues
- in identifying options
- in making choices of action
- in delivering them and
- in monitoring their impact and outcomes.

so…

- Data must be driven by policy needs – whilst maintaining independence.
- Achieving an appropriate balance between relevance and independence is not straightforward especially in situations of resource constraints.

Pre-requisite for evidence based policy is that the data must be trustworthy

- Depends upon
  - the quality of the data
  - the quality of the statistical system
  - the quality of the professional statisticians

But it is not enough that the data are trustworthy they must also be trusted

- otherwise they won't be used
- there will be fights about the data rather than about the issues
- data need to be the currency of public debates

"Trust comes on foot, but leaves on horseback."

Dutch statesman,

Johan Thorbecke

The policy making process

- *Policy making is the process by which governments translate their political vision into programmes and actions to deliver desired changes in the real world*
- Evidence but one input into policy process

Self interest

Values/ beliefs

Campaign promises

Ideology

Judgement

Tradition

Lobbies

Acceptability to public

Experience

Resource constraints

Expert views

All evidence is imperfect

"The absence of excellent evidence does not make evidence-based decision making impossible: what is required is the best evidence available not the best evidence possible"

Muir Gray 1997

" Evidence rarely provides neat and tidy prescriptions to decision makers as to what they should do. Often it generates more questions to be resolved "

Petrosino et al 2001

Evidence sometimes resisted...

" There is nothing a government hates more than to be well-informed: for it makes the process of arriving at decisions much more complicated and difficult. "

John Maynard Keynes

Inconvenient truths

- Governments prefer good news stories
- Bad news stories may be delayed or buried
- They are often too focussed on populism
- The government's horizons can be shorter than those of social researchers!
- They can prefer their own spin to that of the statistician/social researcher

KING'S College LONDON

Sir Gus O'Donnell
(UK Cabinet Secretary)

- "I want [the ONS] to be boring, to put out the plain facts, and nothing but the facts, and on clear, predictable deadlines," he said. It would then be for politicians and government press officers to interpret the figures, he added.

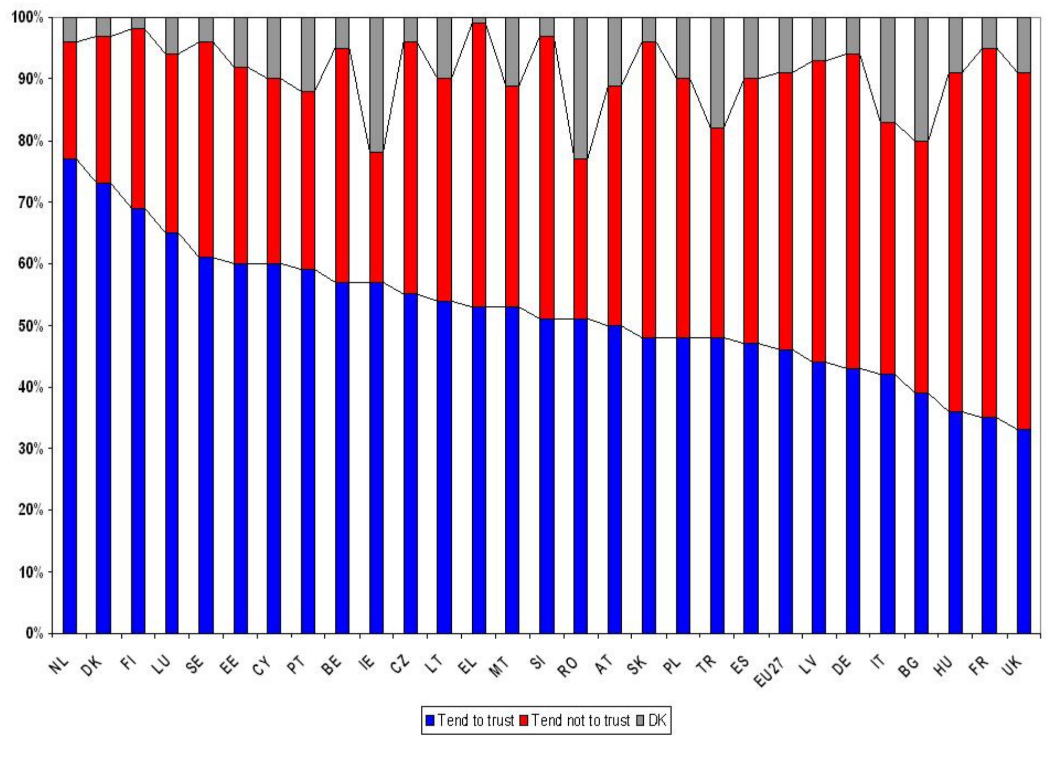
KING'S College LONDON

Response of the Royal Statistical Society

- it is clearly the task of official statisticians to interpret the figures in a statistical context, to facilitate understanding and avoid misunderstanding.
- *The Code of Practice* of the UK Statistics Authority explicitly states that :
  - *Official statistics, accompanied by full and frank commentary, should be readily accessible to all users and that all UK bodies that are responsible for official statistics should prepare and disseminate commentary and analysis that aid interpretation, and provide factual information about the policy or operational context of official statistics.*

Key aspects of building trust

- Autonomy of statistics office
- Statistical legislation
- Existence of an <u>independent</u> statistical board
- Development of codes of conduct
- Breaches of the code identified, investigated and publicised
- Appointment of DG statistics removed from the political process
- Users should be involved in setting the agenda (asking the awkward questions)
- External audits of the statistical processes should be employed
- Audit body should report to Parliament



Counteracting the lack of trust in the UK

- No political interference with the data AND no perception of interference
- Who has access to data prior to its release is critical
- Those who have prior access are identified
- Length of time of prior access is limited
- Data should be released by statisticians and separated from the political spin
- Leakages actively investigated

Challenges to integrity – the rise of performance monitoring

- Performance data can be used
  - to establish 'what works' among policy initiatives
  - to identify well-performing or under-performing institutions and public servants
  - to hold Ministers to account for their stewardship of the public services
- Hence, government is both monitoring the public services, and being monitored by performance indicators.
- Because of government's dual role, performance monitoring must be done with integrity and shielded from undue political influence

Performance indicators – a health warning

- are used as sticks
- can have unintended consequences
- can encourage manipulation
- promote a narrow use of data
- can divert us from addressing the big issues
- need to be carried out with integrity

http://www.rss.org.uk/PDF/PerformanceMonitoring.pdf

Performance monitoring in the international context

- Accountability of governments
- The results matter
- One size fits all – relevance not always obvious
- Distorting effects of measurement
- Validation extra-ordinarily difficult
- National user community often under-developed and under-resourced
- Example – Millennium Development Goals

and the problems …

- Internationally - who sets agenda ?
- Agenda imposed on governments
- Dependence on official data from governments
  - Paucity of data
- Perverse incentives to report in particular ways
- Corruption within statistical systems
  - How do we challenge governments who mislead with data?
- Lack of established professional associations and user communities in poorer countries

What are the challenges to us?

- to collect and report data even if they are uncomfortable for the government of the day
- to address inequities in our societies
- to exercise our responsibilities to use information to improve well- being of global poor

- to build a statistics system that is not only responsive to user needs but also confident and assured - yet not arrogant
- to stimulate collaboration with users who may have expertise greater than ours
- to change the culture so that “service” is not perceived as a derogatory term
- to celebrate the profession of statistics
- to keep our skills up to date by CPD
- to enhance mutual respect across the profession

Priorities for statistics societies - ethical responsibilities

- committees on ethics and other issues of public interest
- development of codes of conduct
- training and mentoring on ethics
- an environment for members to discuss problems
- public statements

Priorities for statistics societies - communication skills

- learning how to tell a story with data
- using the language of policy makers
- understanding users' needs and fitness for purpose
- improving links with responsible journalists (ASA and RSS prizes)
- training journalists
- appointing spokesmen
- producing statistical magazines

Priorities for statistics societies - integration across profession

- strengthening links between academic, research and official statisticians
- facilitating the embedding of statisticians in specialist team
- using the data to answer questions which public employees may not be free to ask
- supporting greater exploitation of existing data
- advocating greater resources for the statistical system

We need leadership

- rooted and committed to the core values of a nation and its people.
- that drives and inspires diverse partners to work collaboratively towards a common objective
- focussed on the excellence of our evidence
- informed by an over-riding commitment to stay grounded and accountable to citizens
- that is bold and dares to dream no little dreams, of how we can build even better nations and, ultimately, a better world.

drawn from Roy Romanow
Founding Chair, The Canadian Index of Wellbeing

14. Planning for Health

Planning for Health

AUSTRALIAN STATISTICAL CONFERENCE

Dr Jim Codde
Director Planning

6th December 2010

Delivering a **Healthy WA**

Government of **Western Australia**
Department of **Health**

SMAHS Health Service Planning Unit

Australia's Health 2010

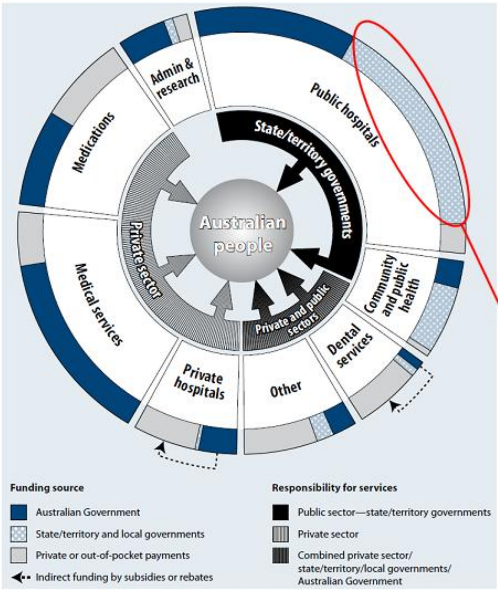
Australia at a glance

- 21.9 million people (June 2009)
- Life expectancy continues to grow
- Fertility rate was 1.97
- 3.7 % aged 80 years and over
- 25% born overseas
- 2.5% Indigenous
- 64% live in capital cities
- 5.5% unemployed
- *Health Expenditure =→ 9.1% GDP ($103b)*

2011 년 1 월 3 일 Slide 2

SMAHS Health Service Planning Unit

Who pays and for what?



- Two thirds of health expenditure is from the Commonwealth, State and Local governments;
- Of this, two thirds is by the Commonwealth government, largely for:
  - Medicare
  - PBS
  - Hospitals

As a State Health planner, we are mainly involved in planning hospital infrastructure and service delivery models (includes EDs).

Source: Australia's Health 2010, AIHW, 2010.

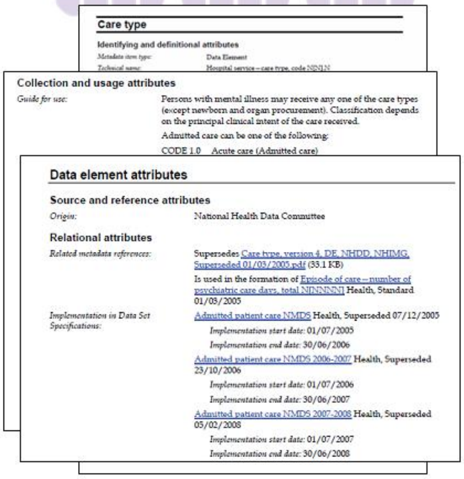
"Study the past if you would define the future."

Confucius

So how do we plan these services?

1. Standardised national health information

- Since at least the early 1980's there has been a nationally agreed definition of health information which is described in National Health Data Dictionary;
- The dictionary also identifies the national agreed minimum data set requirements.
- It provides definitions of the data elements, their codes, guide for use and related attributes.
- Participation is managed through the National Health Information Agreement which is governed by Australian Health Ministers' Advisory Committee (AHMAC).


Care type

Identifying and definitional attributes

Metadata item type: Data Element

Technical name: Hospital service – care type, code N[N].N

Collection and usage attributes

Guide for use: Persons with mental illness may receive any one of the care types (except newborn and organ procurement). Classification depends on the principal clinical intent of the care received.

Admitted care can be one of the following:

CODE 1.0 Acute care (Admitted care)

Data element attributes

Source and reference attributes

Origin: National Health Data Committee

Relational attributes

Related metadata references: Supersedes Care type, version 4, DE, NHDD, NHIMG, Superseded 01/03/2005.pdf (33.1 KB)

Is used in the formation of Episode of care – number of psychiatric care days, total N[NNNN] Health, Standard 01/03/2005

Implementation in Data Set Specifications: Admitted patient care NMDS Health, Superseded 07/12/2005

Implementation start date: 01/07/2005

Implementation end date: 30/06/2006

Admitted patient care NMDS 2006-2007 Health, Superseded 23/10/2006

Implementation start date: 01/07/2006

Implementation end date: 30/06/2007

Admitted patient care NMDS 2007-2008 Health, Superseded 05/02/2008

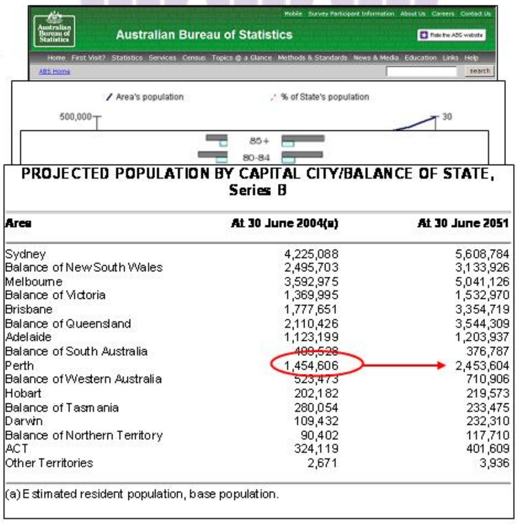
Implementation start date: 01/07/2007

Implementation end date: 30/06/2008

So how do we plan these services?

2. Quality population estimates and projections

- Age and sex specific historical and "current" *Estimated Resident Population* are produced by the ABS at the State and SLA level;
- Similarly, the ABS and WA Dept of Planning provide population projections by age, sex and SLA out to 2051 based on annualized assumptions on fertility, mortality and migration.



PROJECTED POPULATION BY CAPITAL CITY/BALANCE OF STATE, Series B

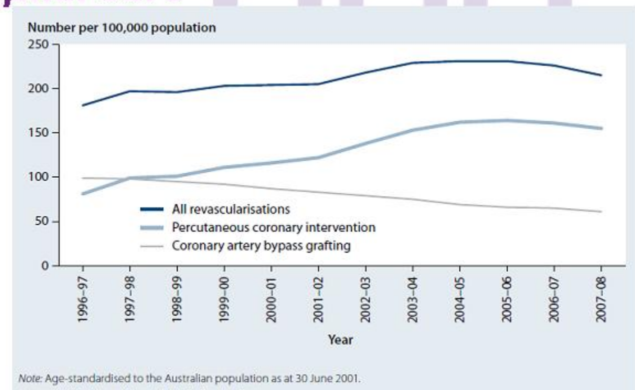
Area	At 30 June 2004(a)	At 30 June 2051
Sydney	4,225,088	5,608,784
Balance of New South Wales	2,495,703	3,133,926
Melbourne	3,592,975	5,041,126
Balance of Victoria	1,369,995	1,532,970
Brisbane	1,777,651	3,354,719
Balance of Queensland	2,110,426	3,544,309
Adelaide	1,123,199	1,203,937
Balance of South Australia	409,528	376,787
Perth	1,454,606	2,453,604
Balance of Western Australia	523,473	710,906
Hobart	202,182	219,573
Balance of Tasmania	280,054	233,475
Darwin	109,432	232,310
Balance of Northern Territory	90,402	117,710
ACT	324,119	401,609
Other Territories	2,671	3,936

(a) Estimated resident population, base population.

So how do we plan these services?

3. Drivers for hospital care

- Population growth;
- Ageing population;
- Age-specific variations;
- Life expectancy;
- Disease prevalence;
- New technology;
- Policy changes.



Note: Age-standardised to the Australian population as at 30 June 2001.
Source: AIHW National Hospital Morbidity Database.

So how do we plan these services?

3. Demand modelling

Medical Records

- Principle Diagnosis
- Comorbidity
- Episode of care
- Procedures
- Age/sex
- Separation mode
- Location of residence
- Hospital

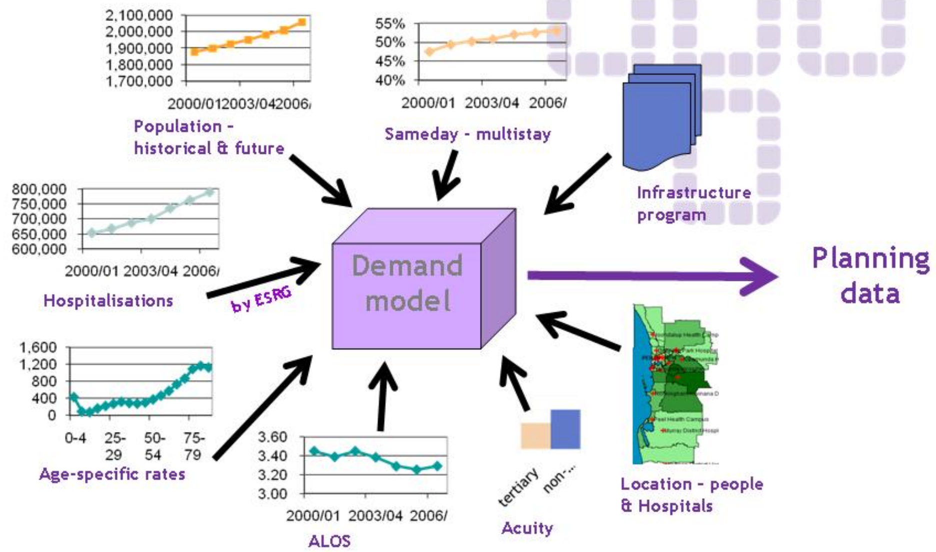
Clinical specialty
@admission /discharge

→ Computer algorithm →

DRG – approx 700
↓
ESRG- approx 128
↓
SRG- approx 44
↓

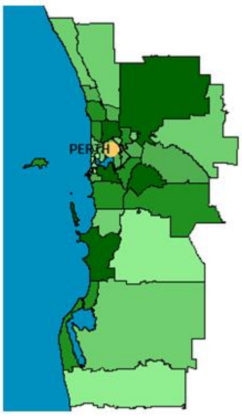
SRGs / ESRGs are used in demand modelling and the development of the CSF & CSP

Demand modelling process

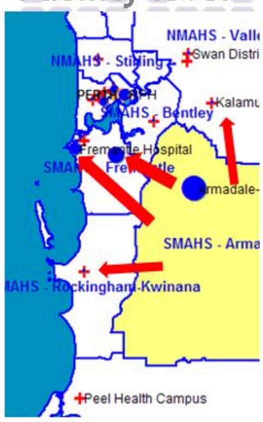


Demand modelling output

Resident level



Facility level



- Conditions
- Sameday - multiday

- Beddays
- Bed category

Projected separations - RPH

hosp_name	RPH
acuity	(All)
mso_name	Surgical
age_group	(All)
bedtype_name	(All)
staytype_name	Multiday

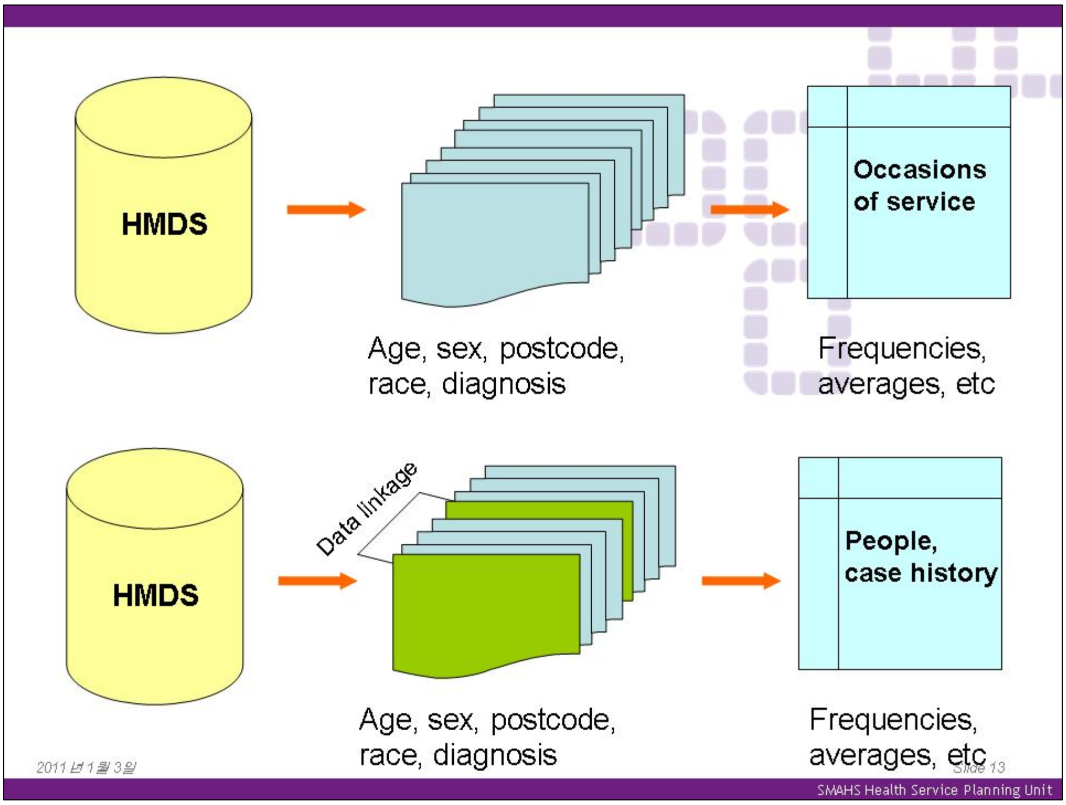
Sum of beds		fy							
srg_name	esrg_name	2007/08	2009/10	2010/11	2011/12	2012/13	2013/14	2014/15	2015/16
17, Breast Surgery	051, Breast Surgery	2	2	2	2	2	3	2	2
17, Breast Surgery Total		2	2	2	2	2	3	2	2
18, Cardiothoracic Surgery	052, Coronary Bypass	5	5	5	5	5	5	5	5
	053, Other Cardiothoracic Surgery	8	8	8	8	8	8	7	7
18, Cardiothoracic Surgery Total		13	13	13	13	13	13	12	12
19, Colorectal Surgery	054, Major S and L Bowel Procs incl Rectal Resection	12	12	12	12	12	12	6	6
	055, Other Colorectal Surgery	2	2	1	1	1	1	1	1
19, Colorectal Surgery Total		14	14	13	14	14	13	7	7
20, Upper GIT Surgery	056, Cholecystectomy	5	4	4	4	4	4	1	1
	058, Other Upper GIT Surgery	6	6	6	5	6	5	4	4
20, Upper GIT Surgery Total		11	10	10	10	10	9	6	6
25, Orthopaedics	070, Wrist and Hand Procedures incl Carpal Tunnel	5	4	4	5	5	4	1	1
	071, Hip & Knee Replacement	4	5	5	5	6	5	9	9
	072, Knee Procedures	1	1	0	1	1	1	0	0
	073, Local Excisn/Removl of Internl Fixn Device Exc Hip/Femur	0	0	0	0	0	0	0	0
	074, Other Orthopaedics - Surgical	35	35	36	37	38	37	25	25
25, Orthopaedics Total		45	45	46	48	49	47	36	37

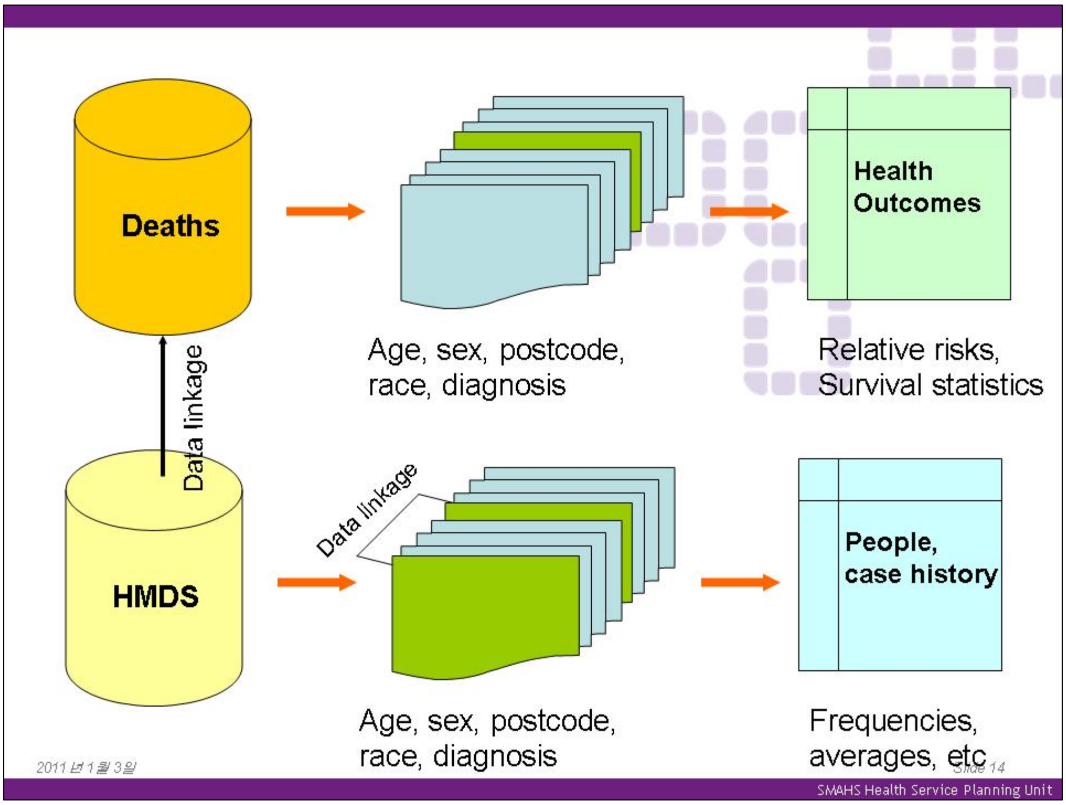
Other applications

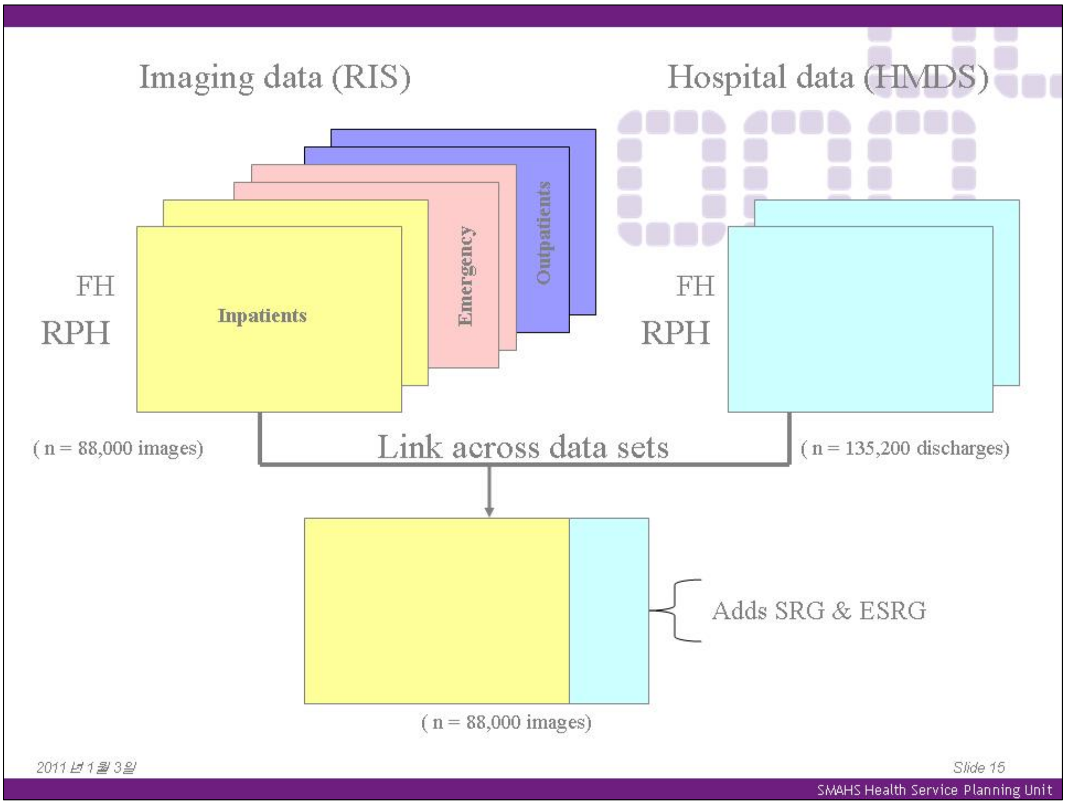
Demand modelled estimates of patient throughput and reason for presentation (ie ESRG) can be used to support a range of purposes that include:

- Determination of hospital size and function
- Clinical infrastructure
  - Theatres
  - Cardiac labs
  - Endoscopy rooms
  - Radiology services
- Workforce requirements
- Activity Based Funding

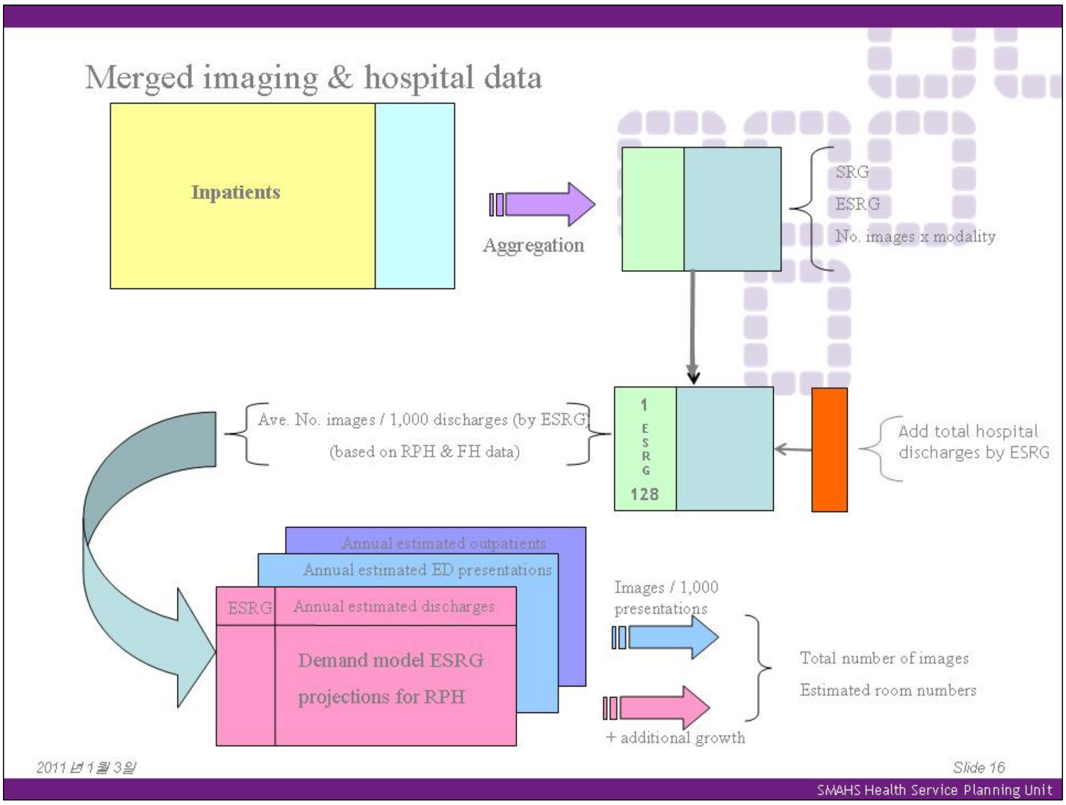
Enhanced through data linkage


HMDS
Occasions of service
Age, sex, postcode, race, diagnosis
Frequencies, averages, etc
HMDS
Data linkage
People, case history
Age, sex, postcode, race, diagnosis
Frequencies, averages, etc
2011년 1월 3일
Slide 13
SMAHS Health Service Planning Unit


Deaths
Health Outcomes
Age, sex, postcode, race, diagnosis
Relative risks, Survival statistics
Data linkage
HMDS
Data linkage
People, case history
Age, sex, postcode, race, diagnosis
Frequencies, averages, etc
2011년 1월 3일
Slide 14
SMAHS Health Service Planning Unit



Imaging data (RIS)
Hospital data (HMDS)
Outpatients
Emergency
Inpatients
FH
RPH
FH
RPH
(n = 88,000 images)
Link across data sets
(n = 135,200 discharges)
Adds SRG & ESRG
(n = 88,000 images)
2011년 1월 3일
Slide 15
SMAHS Health Service Planning Unit



Merged imaging & hospital data
Inpatients
Aggregation
SRG
ESRG
No. images x modality
Ave. No. images / 1,000 discharges (by ESRG)
(based on RPH & FH data)
1
ESRG
128
Add total hospital discharges by ESRG
Annual estimated outpatients
Annual estimated ED presentations
ESRG
Annual estimated discharges
Demand model ESRG projections for RPH
Images / 1,000 presentations
+ additional growth
Total number of images
Estimated room numbers
2011년 1월 3일
Slide 16
SMAHS Health Service Planning Unit

Imaging requirements:

RPH (2014/15)

Examinations by modality

Patient type	BD	CR	CT	DSA	MG	MRI	NM	OT	RF	US	*Total*
Inpatients (n = 43533)	22	23,557	4,910	649	7	2,022	946	2	3,210	3,603	*38,927*
ED (n = 70787)	1	26,655	4,061	11	3	144	51	2	20	1,006	*31,954*
Outpatients (n = 495500)	813	30,012	8,207	120	1,973	6,551	1,292	49	974	7,534	*57,526*
Total	*836*	*80,224*	*17,178*	*779*	*1,983*	*8,717*	*2,290*	*53*	*4,205*	*12,143*	*128,407*

Room requirements by modality

Patient type	BD	CR	CT	DSA	MG	MRI	NM	OT	RF	US	*Total*
Inpatients (n = 43533)	0.0	5.9	1.0	0.6	0.0	0.7	0.9	0.0	1.6	1.8	*13*
ED (n = 70787)	0.0	6.7	0.8	0.0	0.0	0.0	0.1	0.0	0.0	0.5	*8*
Outpatients (n = 495500)	0.2	3.8	1.6	0.1	2.2	2.2	1.3	0.0	0.5	3.8	*16*
Total	*1*	*17*	*4*	*1*	*3*	*3*	*3*	*1*	*3*	*7*	*43*

Assumed annual growth rates and examinations per room

Annual growth in excess of the 3% popu	2%	1%	5%	2%	2.5%	10%	1%	1%	2.5%	1%
Max exams. / 'room'	*4,000*	*4,000*	*5,000*	*1,000*	*900*	*3,000*	*1,000*	*1,000*	*2,000*	*2,000*
Outpatients		8,000								

Conclusion

- State health planning is based on a rich, well defined data that spans socio-demographic characteristics through to health service utilisation information;
- Utilising some of this data, demand modelling has proved a useful tool for informing system reform, health service planning and infrastructure development;
- As with all modelling tools, however, use of sound assumptions, clinical consultation and review of emerging trends are all integral to evidence based health service delivery.


"Sin bravely….We will never have all the facts to make perfect judgement, but with the aid of basic experience we must leap bravely into the future."
Russell R McIntyre
Thank you...
2011년 1월 3일
Slide 19
SMAHS Health Service Planning Unit

15. Weighting and Maximum Likelihood estimation to correct for errors in probabilistically linked datasets

Weighting and Maximum Likelihood estimation to correct for errors in probabilistically linked datasets

James Chipperfield, Glenys Bishop,
Paul Campbell
paul.campbell@abs.gov.au

Australian Bureau of Statistics

Outline

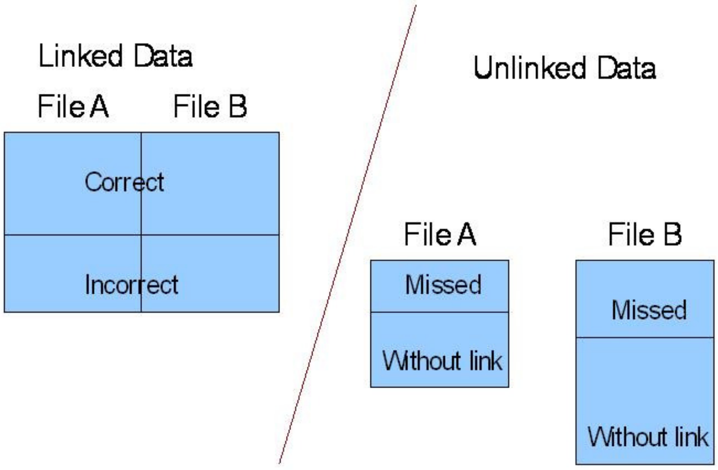
- Structure of probabilistically linked data
- Bias resulting from inexactly linked data
- Weighting to resolve bias
  - Cannot be done naively
- Use a maximum likelihood approach with auxiliary information to adjust

Background

- Data Linking: Linking a population of individuals common to two datasets
  - Resulting dataset has a richer array of information
  - Used to obtain more variables on cross sectional data; or to create longitudinal datasets
- Probabilistic Linking: Data Linking without a unique record identifier
  - Use instead a combination of non-uniquely identifying variables common to both datasets
  - Records which agree on a specified amount of information are deemed a link

File A						File B						Link Weight
First name	Last name	DOB	Sex	Marital Status	Education	First name	Last name	DOB	Sex	Marital Status	Education	
John	Smith	01/01/1978	M	m	1	John	Smith	01/01/1978	M	m	1	19
Archie	Mulliner	03/09/1952	M	d	3	Archie	Mulliner	03/09/1952	M	s	3	17.5
Jean	Jones	25/07/1983	F	s	2	Jean	Jones	25/07/1983	F	s	1	16.5
Sally	Simmons	04/11/1988	F	s	.	Kelly	Simmons	09/11/1988	F	.	3	12
Pongo	Twistleton	29/07/1934	M	m	1	Reginald	Twistleton	29/01/1934	M	.	.	10
Tom	Thomson	30/12/1944	M	m	4	Ron	Johns	30/02/1957	M	.	.	-5

Data Structure



Linking Errors

- Two types of errors
  - Missed Links
  - Incorrect Links
- Errors are related to one another
- These errors result in a biased dataset
  - Missed links are similar to non-response error
    - (but we can't easily identify the non-respondents)

Method 1: Weighting

- In the ideal linked dataset, there are
  - No missed links
  - No incorrect links
  - Records without a link remain unlinked
- Weight to overcome the problem of missed links
- Weight by benchmarking to one of the original (unlinked) datasets

Weighting: Two Simulated Examples

- Two files: X (100 records) and Y (200 records)
  - File X contains variables x = {1,2,3} and z ∈ [0,1]
  - File Y contains variables y = {1,2,3,4} and z ∈ [0,1]
- Continuous variable z is common to both files
  - z is used in linking
- Weight to File X – standardise totals on variable z within categories of variable x

$$w_i = \frac{\sum_{x=i} z_{FileX}}{\sum_{x=i} z_{Linked}}, i = 1, 2, 3$$

- Evaluation: compare weighted and unweighted

Case 1

10 records without a link (random); 40 missed links (high z)

		y = 1	y = 2	y = 3	y = 4
Perfect Linking (90 records)	x = 1	0.111	0.122	0.022	0.000
	x = 2	0.011	0.144	0.100	0.111
	x = 3	0.011	0.022	0.133	0.211
Unweighted linked data with missed links (50 records)	x = 1	0.140	0.180	0.040	0.000
	x = 2	0.000	0.240	0.140	**0.040**
	x = 3	0.000	**0.020**	0.080	0.120
Weighted linked data with missed links (50 records)	x = 1	0.082	**0.105**	**0.023**	0.000
	x = 2	0.000	**0.217**	**0.127**	0.036
	x = 3	0.000	0.037	**0.149**	**0.224**

Case 2

40 records without a link (random); 10 missed links (high z)

		y = 1	y = 2	y = 3	y = 4
Perfect Linking (60 records)	x = 1	0.067	0.150	0.033	0.000
	x = 2	0.017	0.167	0.117	0.133
	x = 3	0.017	0.017	0.133	0.150
Unweighted linked data with missed links (50 records)	x = 1	0.080	0.180	0.040	0.000
	x = 2	0.020	0.200	0.140	**0.140**
	x = 3	0.000	**0.020**	0.080	**0.100**
Weighted linked data with missed links (50 records)	x = 1	**0.057**	**0.129**	**0.029**	0.000
	x = 2	**0.015**	**0.146**	**0.102**	0.102
	x = 3	0.000	0.042	**0.168**	0.210

Case 3

40 records without a link (high z); 10 missed links (random)

		y = 1	y = 2	y = 3	y = 4
Perfect Linking (60 records)	x = 1	0.167	0.183	0.017	0.000
	x = 2	0.000	0.167	0.133	0.100
	x = 3	0.000	0.033	0.050	0.150
Unweighted linked data with missed links (50 records)	x = 1	**0.120**	**0.200**	**0.020**	0.000
	x = 2	0.000	0.200	**0.140**	**0.100**
	x = 3	0.000	**0.040**	**0.040**	**0.140**
Weighted linked data with missed links (50 records)	x = 1	0.080	0.134	0.013	0.000
	x = 2	0.000	**0.172**	0.121	0.086
	x = 3	0.000	0.072	0.072	0.251

Method 2: Maximum Likelihood Adjustment

- Suppose we have a sample of links we know to be correct
- Use this auxiliary information to adjust for linking errors
- Consider ML adjustment in contingency tables and logistic regression

Contingency Tables: Notation

- Variables x and y from Files X and Y
  - x = 1, 2, .., g, .., G
  - y = 1, 2, .., c, .., C
- d ={ (y_i, x_i): $i = 1, \ldots, n_{xy}$ }
- $w_{ic|x}$ = 1 if $y_i = c \mid x_i$

  = 0 otherwise
- $\hat{\pi}_{c|x} = \frac{\sum_i w_{ic|x}}{\sum_C \sum_i w_{iC|x}}$

Notation in Contingency Table

		x					
		1	2	…	g	….	G
y	1	$\sum_i w_{i1|1}$					
	2						
	:						
	c				$\sum_i w_{ic|g}$		
	:						
	C						$\sum_i w_{iC|G}$
Total		$\sum_c \sum_i w_{ic|1}$	..	..	$\sum_c \sum_i w_{ic|g}$	…	…

Adjusting for Incorrect Links

- d* ={d$_i^*$ = (y_i^*,x_i): i = 1, ..., n_x }
  - y_i^* = y_i when link is correct
- $p(y_i, \boldsymbol{x}_i, \delta_i) = p(y_i, \boldsymbol{x}_i; \Pi)p(\boldsymbol{x}_i)p(\delta_i|\boldsymbol{x}_i)$
  - Π = contingency matrix for variables x and y
  - $\delta_i = 1$ if record i on file X is correctly linked
    $= 0$ otherwise
- Assumption: that δ_i is independent from record to record

Adjusting for Incorrect Links (2)

- $\tilde{\pi}_{c|x} = \dfrac{\sum_i \tilde{w}_{ic|x}}{\sum_c \sum_i \tilde{w}_{ic|x}}$
  - $\tilde{w}_{ic|x}$= expectation of $w_{ic|x}$ under perfect linkage given d*
    $= w^*_{ic|x}\, p_{cx} + (1 - p_{cx})\pi_{c|x}$
  - $w^*_{ic|x} = 1$ if $y_i^* = c \mid x_i$
    $= 0$ otherwise
  - p_{cx} = probability of correct linkage, estimated by clerical sample
- Iteratively solve for $\tilde{\pi}_{c|x}$ and $\tilde{w}_{ic|x}$

Logistic Regression

- Model:

$$E(y_i) = v_i \text{ , where } v_i = \frac{1}{1+ e^{\beta' x_i}}$$

- Adjust for incorrect and missed links
  - Technique is similar, model is different

ML Results

- Simulated data
  - Files X and Y each contain 6000 records
    - X contains variable x ~Bernoulli(0.5)
    - Y contains variable y ~Bernoulli(v_i)
      - $v_i = \frac{1}{1+ e^{(-0.5+2.5x_i)}}$
  - Correct links occur with probability *p*
- Assess via
  - Mean Square Error of β

ML Results (2)

		Mean Square Error		
		$p = 0.6$	$p = 0.8$	$p = 1$
Unadjusted	β_0	0.16	0.043	0.0020
	β_1	1.30	0.38	0.0050
ML	β_0	0.013	0.0072	0.0020
	β_1	0.052	0.030	0.0050

Conclusion

- Missed links (along with incorrect links) can lead to an unrepresentative dataset
- Results with simulated data
  - the effects of weighting depends on the structure of linked data
  - ML adjustment is promising
- Weighting is a simple well known process
- ML adjustment may be appropriate for a wider range of linked data
- We need auxiliary information

Future Research

- More testing on empirical data
- Use auxiliary information in weighting

References

- Chipperfield, J., Bishop, G. & Campbell, P. *Maximum Likelihood Estimation for Contingency Tables and Logistic Regression with Incorrectly Linked Data*, Journal of Survey Methodology, in press
- Campbell, P. *Addressing Bias in Linked Data: Weighting to Overcome the Impact of Missed Links on Probabilistically Linked Datasets*

16. Comparing an SLK-based linkage strategy and a name-based linkage strtegy

Australian Government
Australian Institute of Health and Welfare

Comparing an SLK-based linkage strategy and a name-based linkage strategy

Presenter: Andrew Powierski

AUSTRALIAN INSTITUTE OF HEALTH AND WELFARE

Australian Institute of Health and Welfare

Better information and statistics for better health and wellbeing

Outline

1. Purpose of linkage
2. Data
3. Data linkage methods
4. Comparing the links
5. Conclusions

PiAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics for better health and wellbeing

Purpose

- Of Statistical Linkage Keys (SLK)
- Of this SLK-based data linkage
  - link aged care program use data for the Pathways in Aged Care (PIAC) study.
- Of this name-based data linkage
  - To gauge the accuracy of our SLK-based linkage.

PIAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics for better health and wellbeing

Data for the study

- Aged and Community Care Management Information System (ACCMIS):
  - People who used an ACCMIS program from 1 July 2002 to 30 June 2006
  - 415,057 clients
- National Death Index:
  - People who died between 1 July 2003 and 31 December 2006
  - 470,121 records

PIAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics
for better health and wellbeing

SLK-based data linkage: data

- SLK-581
  - 3 letters of surname + 2 letters of given name (5)
  - Date of birth (8)
  - Sex (1)
- Example: Dorothy Windsor 08/06/1921 F

  SLK-581 =

PiAC pathways in aged care

Australian Institute of Health and Welfare

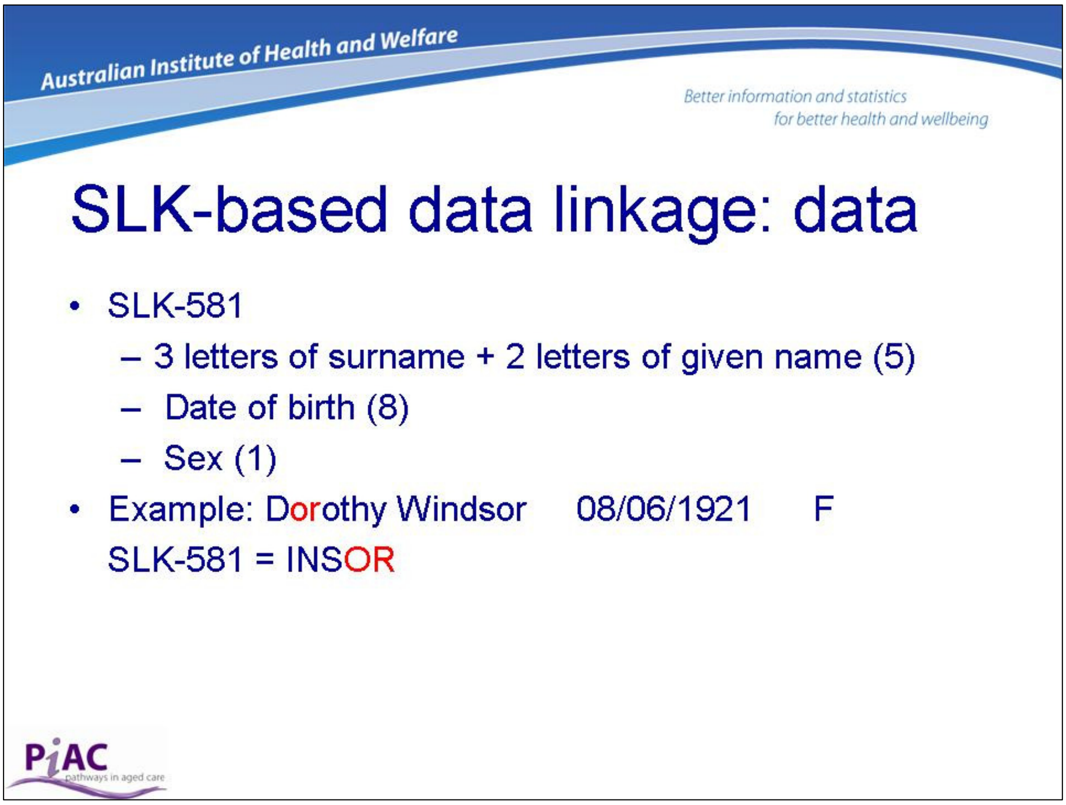
Better information and statistics
for better health and wellbeing

SLK-based data linkage: data

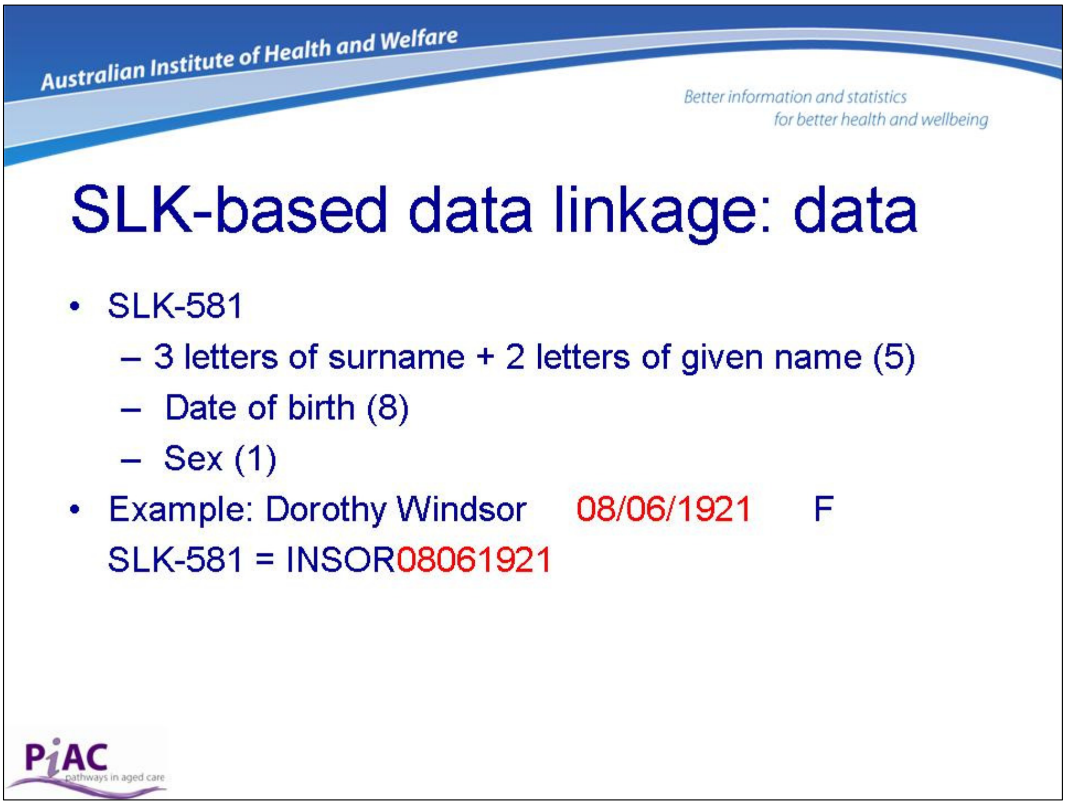
- SLK-581
  - 3 letters of surname + 2 letters of given name (5)
  - Date of birth (8)
  - Sex (1)
- Example: Dorothy Windsor 08/06/1921 F

  SLK-581 = INS

PiAC pathways in aged care



Australian Institute of Health and Welfare
Better information and statistics
for better health and wellbeing
SLK-based data linkage: data
• SLK-581
– 3 letters of surname + 2 letters of given name (5)
– Date of birth (8)
– Sex (1)
• Example: Dorothy Windsor 08/06/1921 F
SLK-581 = INSOR
PiAC
pathways in aged care



Australian Institute of Health and Welfare
Better information and statistics
for better health and wellbeing
SLK-based data linkage: data
• SLK-581
– 3 letters of surname + 2 letters of given name (5)
– Date of birth (8)
– Sex (1)
• Example: Dorothy Windsor 08/06/1921 F
SLK-581 = INSOR08061921
PiAC
pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics
for better health and wellbeing

SLK-based data linkage: data

- SLK-581
  - 3 letters of surname + 2 letters of given name (5)
  - Date of birth (8)
  - Sex (1)
- Example: Dorothy Windsor 08/06/1921 F
  SLK-581 = INSOR08061921F

PiAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics
for better health and wellbeing

SLK-based data linkage: data

- SLK-581
  - 3 letters of surname + 2 letters of given name (5)
  - Date of birth (8)
  - Sex (1)
- Example: Dorothy Windsor 08/06/1921 F
  SLK-581 = INSOR08061921F
- Other shared data items
  - Usual residence postcode(s)
  - Date of death (DOD)

PiAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics for better health and wellbeing

SLK-based linkage: method

- Deterministic linkage
- Match key:
  - SLK-581

PiAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics for better health and wellbeing

SLK-based linkage: method

- Deterministic linkage
- Match key:
  - SLK-581 + pc + DOD

PiAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics for better health and wellbeing

SLK-based linkage: method

- Deterministic linkage
- Match key:
  - SLK-581 + pc + DOD
  - INSOR08061921F

PiAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics for better health and wellbeing

SLK-based linkage: method

- Deterministic linkage
- Match key:
  - SLK-581 + pc + DOD
  - INSOR08061921F435006072005

PiAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics
for better health and wellbeing

SLK-based linkage: method

- Deterministic linkage
- Match key:
  - SLK-581 + pc + DOD
  - INS 0806 F435006072005
- Stepwise deterministic linkage
  - Allows for variation in match keys

PiAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics
for better health and wellbeing

SLK-based linkage: choosing suitable keys

1. *Discriminating power:* 97.5% unique within both datasets
2. *Estimated false match rate* (FMR) ≤ 0.5%.
3. *Trade-off* between additional true and additional false matches: at least 2 to 1.

Sources: Karmel et al. 2010
Karmel & Gibson, 2007

PiAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics for better health and wellbeing

Name-based data linkage: data

- First name and surname
- Date of birth
- Sex
- Other shared data items
  - Usual residence postcode(s)
  - Date of death

PiAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics for better health and wellbeing

Name-based data linkage: method

Probabilistic data linkage

1. Compare records based on blocking variables:
   - Examples: first name, surname, dob, sex, pc, dod
2. Output pairs of all possible links within block and their weights
3. Conduct clerical review (optional)
4. Iterate through steps 1 to 3 based on different blocking variables

PiAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics for better health and wellbeing

Comparing the links

	Name-based links	Name-based non-links
Linked by SLK-based linkage (SLK links)	SLK true links	SLK false links
Not linked by SLK-based linkage (SLK non-links)	SLK missed links	SLK true non-links

PiAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics for better health and wellbeing

SLK true links

Type of link comparison	Outcome	Description	SLK-based links	Name-based links	All links
A	NDI ACCMIS	Same link under name-based and SLK-based linkage	170,127	170,127	170,127
B		Links using SLK-based linkage only	613	..	613
C		Links using name-based linkage only	..	2,460	2,460
D		NDI record links to different ACCMIS record under name-based and SLK –based linkage	147	148	295
E		ACCMIS record links to difference NDI record under name-based and SLK-based linkage	37	38	75
F		Example multiple mixed links	4	3	7
		Total number of links	170,928	172,776	173,577

SLK-based link

Name-based link

PiAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics for better health and wellbeing

SLK false links

Type of link comparison	Outcome	Description	SLK-based links	Name-based links	All links
A	NDI ACCMIS	Same link under name-based and SLK-based linkage	170,127	170,127	170,127
B		Links using SLK-based linkage only	613	..	613
C		Links using name-based linkage only	..	2,460	2,460
D		NDI record links to different ACCMIS record under name-based and SLK –based linkage	147	148	295
E		ACCMIS record links to difference NDI record under name-based and SLK-based linkage	37	38	75
F		Example multiple mixed links	4	3	7
		Total number of links	170,928	172,776	173,577

SLK-based link

Name-based link

PiAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics for better health and wellbeing

SLK false links

- Total of 613
  - ~18% had the same SLK but different names (0.05% of all SLK-based links)
    - Coral Lindsay and Dorothy Windsor
  - ~35% of links were most likely true
    - This result provides evidence that the name-based linkage strategy can miss links.

PiAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics for better health and wellbeing

SLK missed links

Type of link comparison	Outcome	Description	SLK-based links	Name-based links	All links
A	NDI ACCMIS	Same link under name-based and SLK-based linkage	170,127	170,127	170,127
B		Links using SLK-based linkage only	613	..	613
C		Links using name-based linkage only	..	2,460	2,460
D		NDI record links to different ACCMIS record under name-based and SLK –based linkage	147	148	295
E		ACCMIS record links to difference NDI record under name-based and SLK-based linkage	37	38	75
F		Example multiple mixed links	4	3	7
		Total number of links	170,928	172,776	173,577

SLK-based link

Name-based link

PiAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics for better health and wellbeing

SLK missed links

- Total of 2,460
- Links analysed to identify keys that would improve the SLK-based linkage process
- 6 additional keys identified
  - Splitting name information (145 extra links)
  - Splitting day and month of birth / death (246 extra links)

PiAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics for better health and wellbeing

Contested links

Type of link comparison	Outcome	Description	SLK-based links	Name-based links	All links
A	NDI ACCMIS	Same link under name-based and SLK-based linkage	170,127	170,127	170,127
B		Links using SLK-based linkage only	613	..	613
C		Links using name-based linkage only	..	2,460	2,460
D		NDI record links to different ACCMIS record under name-based and SLK –based linkage	147	148	295
E		ACCMIS record links to difference NDI record under name-based and SLK-based linkage	37	38	75
F		Example multiple mixed links	4	3	7
		Total number of links	170,928	172,776	173,577

SLK-based link

Name-based link

PiAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics for better health and wellbeing

Sensitivity and PPV

- Sensitivity: the percentage of all true links that are identified by the SLK-based linkage strategy
- Positive predictive value (PPV): the percentage of SLK-based links that are true links

PiAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics
for better health and wellbeing

Sensitivity and PPV

Table 1: Direct estimates of the PPV and sensitivity of the SLKbased linkage strategy, using name based linkage as the reference standard

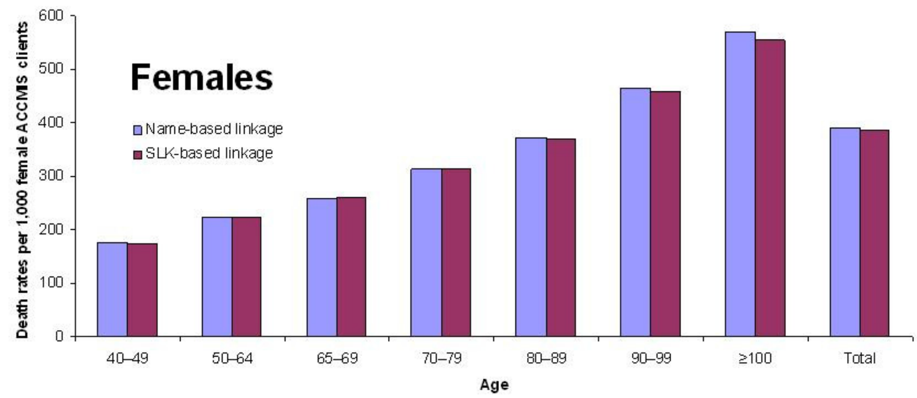
Match Strategy	True links (A)	Additional Links (B)	Missed Links (C)	Total Links (D = A + B)	PPV (A/D)	Sensitivity (A/F)
			Number		Per cent	
Name-based linkage	172,776 **(F)**		..	..	..	
SLK-based linkage	170,127	801	2,649	170,928	99.5	98.5
SLK-581 linkage	152,783	245	19,993	153,028	99.8	88.4

PiAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics
for better health and wellbeing

Utility of SLK-based data: Female death rates by age

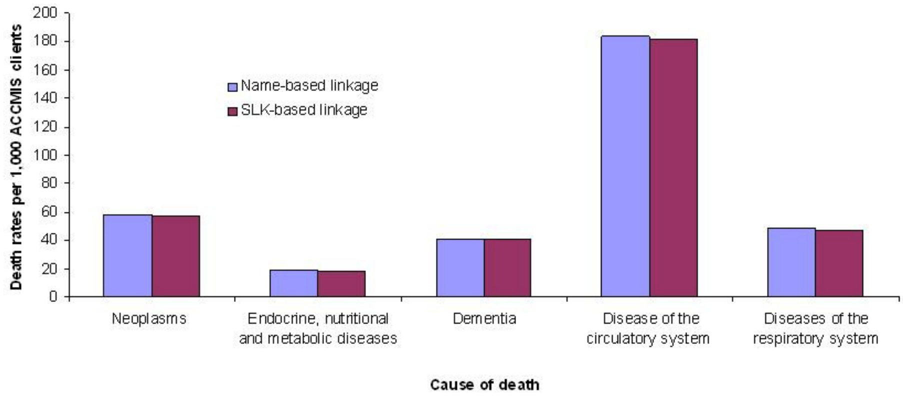


PiAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics for better health and wellbeing

Utility of SLK-based data: Death rates by cause of death



PiAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics for better health and wellbeing

Conclusions

- The stepwise SLK-based strategy was highly effective at identifying matches
- Deterministic matching algorithms can be used to obtain high quality linked data sets for analysis

PiAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics for better health and wellbeing

Publications on linkage methods

- **Event-based record linkage in health and aged care services data: a methodological innovation.** Karmel, R., & Gibson, D. (2007). *BMC Health Services Research, 7*, 154.
- **Empirical aspects of record linkage across multiple data sets using statistical linkage keys: the experience of the PIAC cohort study.** Karmel, R., Anderson, P., Gibson, D., Peut, A., Duckett, S., & Wells, Y. (2010). *BMC Health Services Research, 10:41*
- **Comparing an SLK-based and a name-based data linkage strategy: An investigation into the PIAC linkage.** Powierski, A., Karmel, R., & Anderson, P. (to be published in 2011)

PIAC pathways in aged care

Australian Institute of Health and Welfare

Better information and statistics for better health and wellbeing

Stepwise deterministic linkage:

Estimated FMR when linking 2 datasets using key K deterministically

$$= \frac{\text{Total number of expected chance matches}}{\text{Estimated total number of matches}}$$

$\approx r \times P / \beta\alpha$ where

P	size of the source population for both datasets
r	proportion of P in dataset 1
α	deterministic match rate of datasets using key K
β	number of comparison cells for key K, adjusted for uneven client spread across cells.

PIAC pathways in aged care

Sources: Karmel et al. 2010
Karmel & Gibson, 2007