ISSN 1823-626X

Malaysian Journal of Fundamental and Applied Sciences

available online at http://mjfas.ibnusina.utm.my

Combining multiple survival endpoints within a single statistical analysis

Zakiyah Zain[1] and John Whitehead[2]

[1]*Department of Mathematics and Statistics, School of Quantitative Sciences, College of Arts and Sciences, Universiti Utara Malaysia, UUM Sintok, Kedah, Malaysia.*
[2]*Department of Mathematics and Statistics, Faculty of Science and Technology, Lancaster University, Lancashire, U.K.*

Received 6 December 2012, Revised 19 February 2013, Accepted 22 February 2013, Available online 26 February 2013

ABSTRACT

Multiple endpoints are common in survival data and this scenario complicates the analysis. For example, sets of responses concerned with survival times in a single clinical trial include: time to first cardiac event and time to death from any cause; time to loss of vision in the left eye and time to loss of vision in the right eye; and times from entry to a trial until the first, the second and the third asthma exacerbations. In a clinical trial evaluating the treatment effect of a new drug, often a single statistic is required to measure its overall performance. The cumulative treatment advantage is often measured by the score statistic for each endpoint. The aim of this paper is to develop methodology for combining multiple endpoints within a single statistical analysis that compares the responses of patients treated with a novel treatment with those of control patients treated conventionally. The focus is on interval-censored bivariate survival data, and a real dataset from previous study concerning multiple responses are used for illustration. In this paper we take a direct approach to combining the univariate score statistics for comparing treatments with respect to each survival endpoint. Recurrent events are considered in this investigation and the accuracy of the estimator is evaluated. The combined methodology is accurate, consistent and comparable to the established method of Wei, Lin and Weissfeld.

| Survival analysis | Global test | Score statistics | Recurrent events | Multivariate | Interval-censored |

http://dx.doi.org/10.11113/mjfas.v9n1.82

1. INTRODUCTION

In clinical trials, the main purpose is often to compare efficacy between experimental and control treatments. These treatment comparisons often involve several responses or endpoints, and this situation complicates the analysis. For example, sets of responses concerned with survival times in a single clinical trial include: time to first cardiac event and time to death from any cause; and times from entry to a trial until the first, the second and the third asthma exacerbations. One approach to simplifying the analysis would be to choose one of the survival times as a single primary endpoint. This is not always desirable in cases where the choice would be rather subjective or where the endpoints are of equal interest. A single parameter relating to an overall assessment is often required to give a solid justification of treatment advantage, and so separate analyses of more than one endpoint might not be appropriate.

The cumulative treatment advantage is usually measured by the score statistic for each endpoint. In the case of bivariate survival data, the score statistics can be summed directly, but the variance is now affected by the dependence structure between two endpoints. To estimate the correlation coefficient, an approximate formula for the covariance between the two score statistics is derived for recurrent events data.

This paper discusses the global score test methodology using a new approach of interval-censored data in estimating the correlation. Once the correlation is obtained, the overall treatment effect can be estimated. Accuracy of the method is evaluated and compared with the established method using both real data and simulations.

2. METHODOLOGY

2.1 Global Score Test

In an investigation of a treatment effect θ, an important sample statistic is the cumulative measure of the advantage of the experimental treatment, often denoted by Z. Its companion, denoted by V, indicates the amount of information about θ contained in Z. Statistically termed as the efficient score statistic, and Fisher's information, Z and V, respectively, they can be calculated at any stage of a clinical trial. In survival analysis, the logrank test [1] is one of the most popular methods for testing the equality of two treatment groups. It is routinely used in the analysis of clinical trials comparing the time-to-event distribution of a group of patients randomised to an experimental treatment with that of a control group. When prognostic factors are to be adjusted for, Cox's proportional hazards regression [2],

*Corresponding author. E-mail: zac@uum.edu.my (ZakiyahZain)
Tel:(60)-4-9286967, Fax : (60)-4-9286906

which is a direct generalisation of the logrank test, is commonly employed.

In general, global test methodology can be defined as the use of a combined model to estimate a composite measure of treatment effect concerning multiple outcomes. A global null hypothesis, that the treatment has no effect on any of a number of patient responses, is tested. O'Brien [3] and Pocock, Geller and Tsiatis [4] have combined multiple binary endpoints and reported that global tests may increase the power to detect differences between groups. Global test methodology has been used successfully in major clinical trials involving binary data when multiple outcomes are concerned. It has been accepted in stroke studies, for its ability to yield a single parameter of treatment advantage, which is easily interpreted, as well as for its cost-saving benefit in terms of the trial size.

In particular, use of a global test as a primary analysis for multiple binary outcomes, accompanied by secondary tests of individual outcomes, was implemented in the NINDS t-PA Stroke Trial [5]. Global testing was adopted also for the International Citicoline Trial in acUte Stroke (ICTUS) as reported by Davalos et al. [6] Moreover, Bolland et al. [7] concluded for larger samples that global tests gave accurate type I error rates and satisfactory power, even after adjustment for prognostic factors. Therefore, the global testing approach is attractive for research concerning situations in which two or more time-to-event responses are observed on each individual.

Previous work has successfully determined the correlation between two score statistics arising from binary data or from ordered categorical data [8], but the case of survival data has proved difficult. An existing method for combining two or more survival analyses is the method of Wei, Lin and Weissfeld [9]. Unlike the logrank test, their approach does not directly condition on risk sets and does not reproduce the familiar form of logrank variance.

An earlier approach using the logrank test proved difficult and therefore a new strategy is now proposed. In this new approach, the survival data are summarised within categories and analysed as interval-censored survival data. Using such a formulation, it is possible to determine the correlation, which serves as an accurate approximation to the correlation of the logrank statistics. Correlations between score statistics arising from interval-censored forms of the Cox model are investigated. Once an estimate for the correlation between two score test statistics is available, it has many applications. For example, combined null hypotheses, testing whether a linear combination of effects is equal to zero, and global null hypotheses, testing whether all effects are equal to zero, can be addressed.

2.2 Interval-censored Survival Data

Interval-censored survival data commonly occur in medical or health studies of non-fatal endpoints requiring regular follow-ups or inspections. Consider the case of tumour recurrence where no recurrence had been observed at a three months examination, but one was detected at a six months check-up. It is known that the event time is greater than three months and less than or equal to six months: $3 < T \leq 6$. Another common scenario of interval-censored data is present when continuous survival times are grouped into defined intervals prior to analysis. In practice, survival data are often observed to the nearest time unit: day, month or year, and hence the analyses are generally based on interval-censored data. Consequently, it is natural to consider the underlying survival variables as discrete in developing methods for their analysis. A survival text by Sun [10] provides a comprehensive coverage of the topic of interval-censored survival data.

2.3 Bivariate Survival Data

Bivariate survival data involves two endpoints which cannot be assumed to be independent, and one of the main interests in the analysis of bivariate survival data is the measure of dependence or association of these two variables. The complexity of studies concerning such correlated times-to-event which may involve multiple endpoints on the same subject, requires methods to take into accountthe correlation between multiple endpoints. For such data, the correlation between two score statistics can be used to obtain an overall treatment efficacy.

In dealing with correlated survival outcomes in cross-over trials, fixed effects models can be applied by fitting Cox's proportional hazards regression model stratified by subject. However, in parallel group trials where patients are randomized to experimental and control treatments, these methods fail. To overcome this difficulty, recourse can be made to one of two methods, namely marginal and frailty modelling. Marginal modelling involves fitting data to Cox's regression model without any assumption of correlation, and then adjusting the estimated variance of the coefficients. A frailty model is a random effects model for event time data where subject effects are modelled as random variables; a good description is given by Hougaard [11].

2.4 Estimating the Correlation

The bivariate survival data are first categorized into multiple intervals, and then the covariance is obtained directly from two survival endpoints, say T_1 and T_2 for each subject. Based on the Cox's model of proportional hazards assumptions, the score statistic Z and Fisher's information, V are derived from complementary log log approach. Conditioning on successive risk sets, the covariance between two score statistics, $Cov(Z_1, Z_2)$ is obtained by the summation of covariances from each pair of intervals, and is denoted by C_{12}.

The derivation of the covariance involves two parts: the marginal given by interval of individual event and the paired intervals of both events. To fully describe this, the following notations are necessary, where:

r_{1i} = no. of patients at risk of event 1 at the end of interval *i*,
o_{1i} = no of patients who had event 1 at the end of interval *i*,
r_{2j} = no. of patients at risk of event 2 at the end of interval *j*,
o_{2j} = no of patients who had event 2 at the end of interval *j*,
$q_{1i} = -log\ (1-o_{1i}/r_{1i})$ and similarly $q_{2j} = -log\ (1-o_{2j}/r_{2j})$.

The probability of occurrence for both events can be approximated by $\hat{p}_{(12),(ij)} = o_{(12),(ij)} / r_{(12),(ij)}$, and similarly for individual event, $\hat{p}_{(1\bullet),(ij)} = o_{(1\bullet),(ij)} / r_{(12),(ij)}$ and $\hat{p}_{(\bullet 2),(ij)} = o_{(\bullet 2),(ij)} / r_{(12),(ij)}$, where:

$o_{(12),(ij)}$ = no. of patients who had event 1 during interval *i* and also had event 2 during interval *j*,

$r_{(12),(ij)}$ = no. of patients at risk of event1 at the end of interval *i* and also at risk of event 2 at the end of interval *j*,

$o_{(1\bullet),(ij)}$ = no. of patients who had event 1 during interval *i* and also at risk of event 2 (but did not have event 2) during interval *j*, and

$o_{(\bullet 2),(ij)}$ = no. of patients who had event 2 during interval *j* and also at risk of event 1 (but did not have event 1) during interval *i*.

In our proposed method called ZW (Zain & Whitehead), the covariance of the two score statistics is estimated by

$$Cov(Z_{1i}, Z_{2j}) = \frac{q_{1i}q_{2j}}{o_{1i}o_{2j}}\left\{(r_{1iE}r_{2jE}r_{(12),(ij)C} + r_{1iC}r_{2jC}r_{(12),(ij)E})(\hat{p}_{(12),(ij)} - \hat{p}_{(1\bullet),(ij)}\hat{p}_{(\bullet 2),(ij)})\right\}.$$

Plugging the estimates of probability of occurrence of both events, the covariance estimator, denoted by $C_{12(ij)}$ is written as

$$C_{12(ij)} = \frac{q_{1i}q_{2j}}{o_{1i}o_{2j}r_{(12),(ij)}^2}\left\{(r_{1iE}r_{2jE}r_{(12),(ij)C} + r_{1iC}r_{2jC}r_{(12),(ij)E})(r_{(12),(ij)}o_{(12),(ij)} - o_{(1\bullet),(ij)}o_{(\bullet 2),(ij)})\right\}.$$

The theory states that for very large samples $n \to \infty$, $V_1 \to \text{var}(Z_1)$, $V_2 \to \text{var}(Z_2)$, and the estimate $C_{12} \to \text{cov}(Z_1, Z_2)$. The covariance of each pair of intervals is summed to give the total covariance:

$$C_{12} = \sum_{i=1}^{u}\sum_{j=1}^{v} C_{12(ij)}.$$

The correlation ρ between these two score statistics is expressed as $\rho = \text{cov}(Z_1, Z_2) / \sqrt{V_1 V_2}.$, and can be estimated by

$$\hat{\rho} = C_{12} / \sqrt{V_1 V_2}.$$

Under the null hypothesis that the two treatment groups have identical survival experience, the experimental has zero treatment effect and the proportional hazards assumption is true. There exists a common treatment advantage, $\theta_1 = \theta_2 = \theta$ and under H_0: $\theta = 0$; hence p-values are always valid. However, under the alternative, H_1: $\theta_1 = \theta_2 = \theta$, but $\theta \neq 0$. The logrank test is efficient in detecting such a proportional hazards alternative. When the assumption of equal treatment effect is met, the complex multivariate problem of analyzing the multiple endpoints is simplified to the univariate problem of comparing the common effect across the treatments. Even if the equality assumption is not met, the power should be good if the spread of θ is reasonably small. The estimated common treatment advantage is given by $\hat{\theta} = w_1\hat{\theta}_1 + w_2\hat{\theta}_2$ where *w* is some weighting with subscripts 1 and 2 for the 1st and 2nd events respectively, and $w_1 + w_2 = 1$. Since the endpoints are not independent, an optimal weighting [12] is employed as it yields the smallest variance out of all weighted averages of θ_1 and θ_2.

2.5 Analysis of Recurrent Events

To illustrate for recurrent events, bladder cancer data sets based on a study conducted by the Veterans Administration Cooperative Urological Research Group is used. The complete data is listed in Wei, Lin, and Weissfeld [7]. The study comprises 86 patients with superficial bladder tumours, which were removed transurethrally when the patients entered the study; 48 were randomized into the placebo group (control), and 38 were randomized into the thiotepa group (experimental). The majority of patients experienced multiple recurrences of tumours during the study, and new tumours were removed at each visit. The original data set contains the first four recurrences of the tumour for each patient, and each recurrence time was measured from the patient's entry time into the study. However, our analysis setting is limited to the first and second recurrences, and only one covariate that is treatment group. Out of the 86 patients, 47 patients have only one tumour recurrence, while 29 patients have two recurrences.

Our method gives a correlation estimate of 0.619: compared to 0.643 as computed from the Wei Lin and Weissfeld (WLW) method. The estimates of overall treatment effect using ZW and WLW are 0.411 (s.e. 0.284, p = 0.169) and 0.401 (s.e. 0.232, p = 0.148) respectively. The sum of score statistics derived from this interval-censored data are similar to the logrank statistics and their corresponding Fisher's information in parentheses; Z_1 = 4.47 (4.09), Z_2 = 3.94 (3.83), V_1 = 11.82 (10.99) and V_2 = 7.35 (7.06). These results show that our method is comparable to WLW, but simulation study is the confirmatory.

3. SIMULATIONS

To evaluate accuracy of each method, simulation with 20,000 replications is conducted for bivariate case of recurrent events. The estimated correlation values of each method are compared against the correlation values derived from its own samples. A fixed sample size, n = 1,000 are generated from a random uniform distribution *U(0,1)* and randomized equally to control, *C*, and experimental, *E*. Inputting the hypothetical values of $\lambda_{E(0)} = 0.004$ and $\lambda_{C(0)} = 0.006$, the survival times, T_1 and T_2 (days) are generated from an exponential distribution, $T_{im} \sim EXP(\lambda_G \exp(s_i))$, where s_i is a subject effect for patient*i*, following a normal distribution $N(0, \sigma^2)$. The standard deviation σ of the subject effect is set to be $d(\log\lambda_C - \log\lambda_E)$, where *d* is a constant

multiplier chosen to impose varying degrees of correlation: setting d = 1, 5 and 10, creates low, medium and high correlations, respectively.

The Cox's PH requires non-informative censoring such that the censoring is independent of the survival times. Therefore, assuming equal hazards, $\lambda_C = \lambda_E = \lambda$, an overall censoring variable, $C_i \sim EXP(2\lambda y)$ is applied to the whole data set (both patients on C and E), where $y = x/\{2(1-x)\}$ and x is the censoring proportion. It is to be recalled that, in general, patient i is censored for an event when $C_i < T_i$ for that event. Based on the useful expression, $V = bn$, and also $V = \{(u_\alpha + u_\beta)/\theta\}^2$, the θ at which a given power is achieved can be determined by fixing the sample size n, and finding the constant b. The value of b is found to converge satisfactorily when n = 1 million. The type I error rate is targeted at 5% level (2-sided) and the power is aimed at 90%; hence $(u_{\alpha/2} + u_\beta)^2 = 10.51$. The effect of increasing the standard deviation of the subject effect, σ is also investigated by varying the values of d. Each data set is generated based on the variables set for each value of σ, on each hypothesis. The score statistics, Fisher's information and covariance for each interval are computed to yield the global score statistics and covariance estimator, C_{12}. All simulation runs are replicated 10,000 times under each hypothesis, from which the average values are taken to be the best estimates.

4. RESULTS & DISCUSSION

4.1 Simulation Results

Key performance measures considered are type I error, power and correlation ratio. For correlation ratio, an ideal situation is when the estimated correlation, $\rho_{(est)}$ is exactly the same as $\rho_{(sample)}$, the correlation observed from its own samples of 10,000. The estimate of the covariance between two score statistics, C_{12}, is calculated from each replicate simulation and similarly for the correlation estimate $\hat{\rho}$. The average value of $\hat{\rho}$ from the 10,000 replicates, gives the best estimate, $\rho_{(est)}$. Since the "true" correlation is unknown, it is assumed that the correlation observed from its own samples of N, denoted by $\rho_{(sample)}$ gives the true correlation asymptotically. With N replicates of samples of size n, the sample covariance, $Cov(Z_1, Z_2)$ can be obtained from the expression:

$$\mathrm{cov}(Z_1, Z_2) = \left(\sum Z_1 Z_2 - \left(\left(\sum Z_1 \sum Z_2\right)/N\right)/(N-1)\right).$$

The correlation derived from the sample covariance is given by $\rho_{(sample)} = \mathrm{cov}(Z_1, Z_2)/\sqrt{\mathrm{var}(Z_1)\mathrm{var}Z_2}$. Therefore, the correlation ratio of both estimates, $\rho_{(est)}/\rho_{(sample)}$, will be compared in investigating the properties of the correlation estimator and in evaluating the accuracy of this method.

Table 1 shows that both methods give type I error rates within the 95% probability interval (0.022, 0.028). The power of our method is comparable to that of WLW. Both methods give power of equal or more than 0.89 (1-β = 0.90 ± 0.01) at low correlation (d = 1), but when a higher correlation is imposed by increasing the subject effect, the power reduces accordingly for both methods.

With respect to the correlation estimates, Fig. 1 shows accurate estimation of the correlation for WLW, with $y = 0.9964x$, while our method shows an underestimation of about 2%. Theoretical equality of the correlation between two estimates of treatment advantage and that between two score statistics is proven accordingly.

Table 1. Simulation results for ZW and WLW methods under the null and alternative hypotheses.

WLW: H_0										
d	θ	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}$	α_1	α_2	α_{12}	N/A	$\rho_{(est)}$	$\rho_{(sample)}$
1	0.000	0.000	0.001	0.000	0.026	0.025	0.026	N/A	0.556	0.557
5	0.000	0.001	0.000	0.001	0.027	0.025	**0.028**	N/A	0.745	0.753
10	0.000	0.001	0.001	0.001	0.025	0.025	0.026	N/A	0.856	0.856
WLW: H_1										
d	θ	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}$	$1-\beta_1$	$1-\beta_2$	$1-\beta_{12}$	TP	$\rho_{(est)}$	$\rho_{(sample)}$
1	0.293	0.283	**0.437**	0.318	0.87	**0.96**	0.95	0.94	0.552	0.552
5	0.311	0.192	0.240	0.200	0.54	0.57	0.58	0.55	0.744	0.750
10	0.317	0.117	0.131	0.118	0.23	0.23	0.24	0.22	0.851	0.853
ZW: H_0 @ 5 intervals										
d	θ	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}$	α_1	α_2	α_{12}	N/A	$\rho_{(est)}$	$\rho_{(sample)}$
1	0.000	0.000	0.001	0.000	0.026	0.025	0.027	N/A	0.539	0.557
5	0.000	0.001	0.001	0.001	**0.028**	0.025	**0.028**	N/A	0.730	0.751
10	0.000	0.001	0.001	0.001	0.026	0.025	0.026	N/A	0.848	0.855
ZW: H_1 @ 5 intervals										
d	θ	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}$	$1-\beta_1$	$1-\beta_2$	$1-\beta_{12}$	TP	$\rho_{(est)}$	$\rho_{(sample)}$
1	0.293	0.282	**0.430**	0.317	0.86	**0.96**	0.95	0.94	0.543	0.555
5	0.311	0.192	0.238	0.200	0.53	0.57	0.58	0.55	0.730	0.749
10	0.317	0.118	0.131	0.120	0.24	0.24	0.24	0.23	0.848	0.856

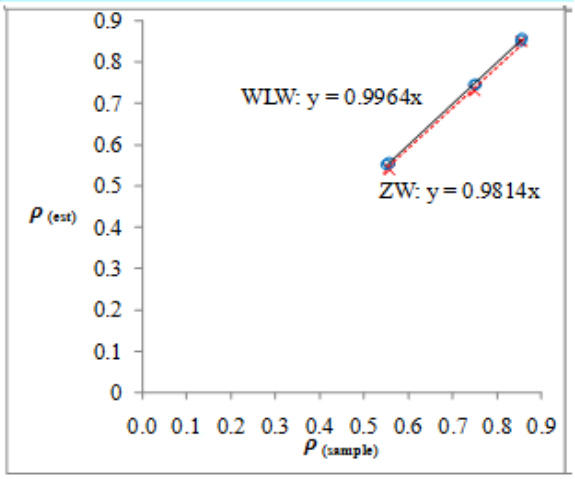

Fig. 1. Correlation ratios for ZW and WLW methods

4.2 Discussion

ZW provides a conceptually straightforward approach to the analysis of general multivariate survival data. The benefits of our method are (i) ease of use: simple computation and (ii) good interpretability: straightforward derivation based on marginal analyses, unlike WLW which is based on non-standard adjusted statistics. By design, ZW is capable of analyzing naturally interval-censored data whereas WLW was not intended to cater for such data. Meanwhile the disadvantages are that it requires categorization into intervals, and consequently might lose a little power. Despite their technical differences, extensive simulations show that our new method is accurate, consistent and comparable to WLW in all scenarios investigated.

The much emphasized issue of overestimation by WLW [13], is not solely due to the risk set definition, but rather an inevitable scenario when using total time convention. Kelly and Lim [13] also commented that the within subject correlation was not satisfactorily accounted for by employing the robust variance, but the reason is unknown. It was suggested that frailty model might be of a better choice and this topic could be further explored.

The methodology for survival analysis of recurrent events has been applied in many diverse fields: numerous examples from medicine, manufacturing and the social sciences are given by Nelson [14]. Others include biostatistics [15], marketing [16], and sports [17] and even in political science [18].

5. CONCLUSION

It is concluded that our method is accurate, consistent and comparable to the competitor. Workable for practical application on real data and comparable to WLW, our proven method ought to contribute a new alternative method of analyzing correlated survival data. Areas for further development include adjustment for combining survival and binary data, and implementation for multiple or sequential methods. Apart from medicine, other potential fields of application include manufacturing, engineering and social sciences.

ACKNOWLEDGEMENT

The authors would like to thank the Department of Mathematics and Statistics of Lancaster University and the School of Quantitative Sciences, Universiti Utara Malaysia.

REFERENCES

[1] R. Peto, and J. Peto, Journal of the Royal Statistical Society Series A-General, 135 (1972) 185-207.
[2] D. R. Cox, Journal of the Royal Statistical Society Series B-Statistical Methodology, 74 (1972) 187-220.
[3] P. O'Brien, Biometrics, 40 (1984) 1079-1087.
[4] S. J.Pocock, N. L.Geller, and A. A. Tsiatis, Biometrics, 43 (1987) 487-498.
[5] B. C. Tilley, J.Marler, N. L. Geller, et al., Stroke, 27 (1996) 2136-2142.
[6] A. Davalos, J.Alvarez-Sabín, J.Castillo, et al.,www.thelancet.com, Published online June 11, 2012 DOI:10.1016/S0140-6736(12)60813-7
[7] K.Bolland, J. Whitehead, E.Cobo, and J. J. Secades, Pharmaceutical Statistics, 8 (2009) 136-149.
[8] J. Whitehead, M.Branson, and S. Todd, Statistics in Medicine (2010)521-532.
[9] L.J.Wei,D.Y. Lin, and L. Weissfeld, Journal of the American Statistical Association, 84 (1989) 1065-73.
[10] J. Sun, The statistical analysis of interval-censored failure time data, Springer, USA,2006.
[11] P. Hougaard, Analysis of multivariate survival data, Springer,New York,2000.
[12] L. J.Wei, and W. E. Johnson, Biometrika, 72 (1985) 359-364.
[13] P. J. Kelly, and L. L.Y. Lim, Statistics in Medicine, 19 (2000) 13–33.

[14] W. B. Nelson, Recurrent Events Analysis for Product Repairs, Disease Recurrences, and Other Applications, ASA, Philadelphia, 2003.
[15] B.Genser, and K. D. Wernecke, Biometrical Journal, 47 (2005) 388-401.
[16] G. E. Bijwaard, P. H.Franses, and R. Paap, Journal of Business & Economic Statistics, 24 (2006) 487-502.
[17] E.Gutierrez, S.Lozano, and J. R. Gonzalez, IMA Journal of Management Mathematics, 22 (2011) 115-128.
[18] J. M.Box-Steffensmeier, and C. Zorn, Journal of Politics, 64 (2002) 1069-1094