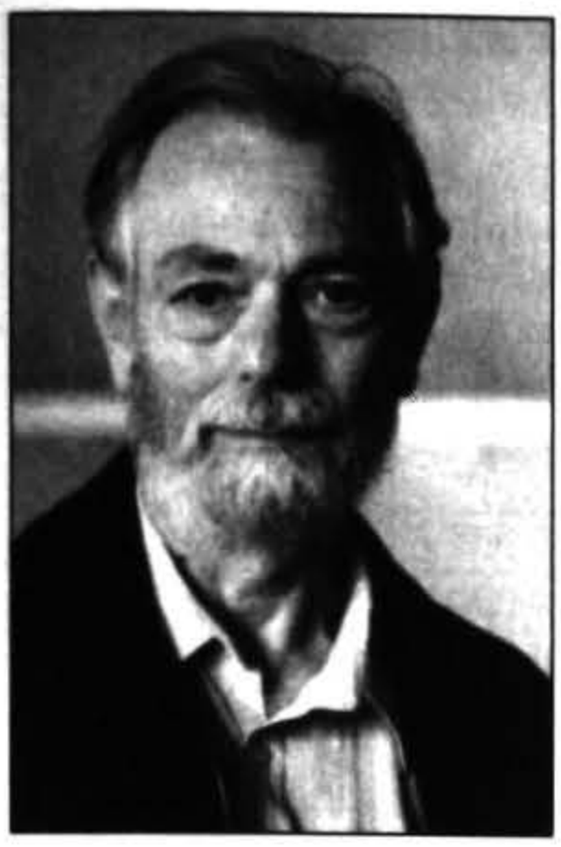


SURVIVAL ANALYSIS OF TRANSITIONS FROM BENEFIT TO WORK USING ADMINISTRATIVE DATA

Terry Moore

Statistical Methods,
Statistics New Zealand, Wellington

Abstract

What factors affect the probability that a person makes a transition from benefit to employment? What is the effect of those factors? Given information such as age, sex, most recent occupation and industry, can we estimate the probability of such a transition? We applied the proportional hazards model to Linked Employer-Employee Data (LEED) to answer those questions. The anonymous longitudinal administrative data is from Inland Revenue and is based on monthly returns. Our principal finding was that, of the limited variables available, age and sex have the most significant impact, and that the difference between sexes is greatest in under-35-year-olds. We also found differences by industry and occupation, as well as some regional differences and time effects.

Introduction

Survival analysis is the study of data models in which individuals experience a change of state (a transition) at a random time. For example, a person may change from being unemployed to employed. In this study we were interested in transitions from benefit to work, and the effects of various explanatory variables, such as age and sex, on the probability of such transitions.

Many studies of labour market transitions have been carried out overseas using survival analysis methods. However, most studies of labour market transitions in New Zealand, e.g. Hyslop *et al.* (2004), have not used this approach. An exception is Moore (2004), who used panel data from the New Zealand Household Labour Force Survey.

Some previous studies have been methodological. For example, Beamonte and Bermúdez (2003) studied a Bayesian additive model and applied it to transitions to first-time employment for graduates. Other studies were designed to address specific research questions. For example, Knut and Zhang (2003) investigated the effects of unemployment compensation on the duration of unemployment, while Gallo *et al.* (2006) investigated the impact of job loss in older workers on the incidences of heart attacks and strokes.

Most previous studies have been based on survey data, e.g. Caroll (2006), who incorporated search theory into his methodology. However, a few studies have used administrative data. For example, Lüdemann *et al.* (2005) studied the length of unemployment periods in West Germany, but, unlike in the present paper, they used a quantile regression model.

The use of such administrative data has both advantages and disadvantages. A complete census of the population has no sampling error nor negligible bias. However, the variables available are usually limited to those collected for a particular administrative purpose.

Linked Employer-Employee Data (LEED)

Linked Employer-Employee Data (LEED) consists of monthly information from Inland Revenue on all taxpayers from 1999 which has been protected for confidentiality. It includes income tested benefits, but not working for families or accommodation allowance. Because it is longitudinal data, we may gain more insight by following the same people over time than would be afforded by cross-sectional data. In particular, we can calculate gross flows between states (e.g. from benefit to wages and salary) rather than net flows. For example, we may see movement in both directions between two categories rather than just the difference between the two movements.

The data includes some imputed variables. For example, the sex of the taxpayer is not available, but the title and name are available on the raw dataset. From this information, the sex is known with a fair degree of certainty. The data also includes information linked from the Business Frame (the list of businesses from which Statistics New Zealand takes samples for business surveys), e.g. the industry of the employer and the number of employees for each business.

Survival Analysis

Survival analysis consists of techniques to assess what factors affect the length of time an individual spends in a category before making a transition to another category, and estimating the size of those effects.

Ideally, we would like to estimate the time that people spend on an unemployment benefit before returning to work. Unfortunately, LEED only tells us if income was derived from an income tested benefit, but the kind of benefit is not specified. It might be possible to gather strength from other data sources, e.g. by random imputation of unemployment based on knowledge of the proportions of unemployed in various age-sex groupings, but we have not done so.

We say that an individual receiving a benefit is at risk of a transition from benefit to work. We may define the term 'at risk' for other transitions in a similar way.

In this study we considered transitions from benefit to paid work. Other competing risks, e.g. to pension, accident compensation or student, were treated as censored data and not included.

Data Analysis

As LEED is so large we worked with a random sample of 10,000 people who had records from December 1999 or before. We then followed those people to the present time, resulting in a total of about 700,000 records.

Anyone who changed status within a given month would have had more than one source of income, and the exact time of the transition was not always known. For this reason we defined the state as the one that had the greatest income (not counting lump-sum payments) during the month. For example, a person was regarded as receiving a benefit during the month if the income tested benefit gave the greatest income. Thus, very short spells of employment were not counted.

Flows Between States

Table 1 shows the gross flows between the proportions of people between states (accident compensation, income tested benefit, pension, paid parental leave, student, and wages and salary) for November and December 2004. The proportions are fairly stable over time, except for the small categories.

Survival Models

The survival function in a state is the probability of exiting that state as a function of time. 'Time' may be defined as time from the start of the study, time in the state or, perhaps, age. In continuous time, the hazard function is the probability density for exiting the state at a given time, conditional on survival to that time. In discrete time, the hazard function is the probability of exiting the state at a given time, conditional on survival to the previous time. (The seemingly negative terminology is derived from the reliability and mortality literature in which the change of state is usually failure or death.)

Table 1: Flows between Principle Income States.

Status Nov 2004	Status Dec 2004	sex		
		F	M	total
ACC	ACC	0.0059	0.0076	0.0135
	W&S	0.0001	0.0001	0.0003
	total	0.0061	0.0077	0.0138
BEN	BEN	0.0638	0.0301	0.0939
	W&S	0.0001	0.0001	0.0003
	total	0.0639	0.0302	0.0941
PEN	PEN	0.1189	0.0788	0.1977
	total	0.1189	0.0788	0.1977
PPL	PPL	0.0004	.	0.0004
	total	0.0004	.	0.0004
STU	STU	0.0018	0.0013	0.0031
	W&S	.	0.0001	0.0001
	total	0.0018	0.0015	0.0032
W&S	BEN	.	0.0001	0.0001
	W&S	0.3266	0.3640	0.6906
	total	0.3266	0.3641	0.6908
total	total	0.5177	0.4823	1.0000

Key:
ACC = Accident Compensation
W&S = Wage and salary
BEN = Benefit
PEN = Pension
STU = Student
PPL = Paid parental leave

For example, if failures occur randomly then the number of failures in a given time interval follows the Poisson distribution, and the time to failure follows an exponential distribution. For this distribution the hazard function is constant and the cumulative hazard function is proportional to time. Popular models that allow for varying hazard functions are the Gompertz distribution, often used in mortality studies, and the Weibull distribution, often used in reliability theory.

We use the proportional hazards model, which assumes an unknown baseline hazard function and finds how the explanatory variables affect this. This is a semi-parametric model – a fully parametric model for the effects of the explanatory variables and a non-parametric model for the baseline hazard function. For those variables that have a significant effect we give the hazard ratio. This is a factor by which the baseline hazard function is multiplied for a unit change in that particular variable.

Survival Analysis Theory

This section contains a brief summary of the principal formulae of survival theory. Readers more interested in the results may omit it.

Let the survival function be $S(t) = P(T > t)$ where T is the time at which the state is exited. Let the hazard function for exiting the state be $h(t) = \frac{f(t)}{S(t)}$ where $f(t)$ is the probability density for exit. The cumulative hazard function is obtained by integrating the hazard function with respect to time. Its importance is in its relationship to the survival function. We have $H(t) = \int_0^t h(u)du = -\ln(S(t)$, $h(t) = \frac{dH(t)}{dt}$ and $S(t) = \exp(-H(t))$ where $H(t)$ is the cumulative hazard function.

If no distributional assumptions are made then non-parametric methods may be used to estimate the survival function. Typically, these are only about 60 percent efficient relative to correctly specified parametric models, but are a protection against an invalid model, Therneau and Grambsch (2000). The Kaplan-Meier estimator produces the survival function directly, from which we may find the cumulative hazard function: $\hat{S}(t) = \prod_{j:t_j \le t} \left(1 - \frac{\#\text{failures at time } t_j}{\#\text{at risk at time } t_j}\right)$, $\hat{H}(t) = -\ln(\hat{S}(t))$. The Breslow estimator works the other way round, starting from the Nelson-Aalen estimator, Therneau and Grambsch (2000), of the hazard function: $\hat{H}(t) = \sum_{j:t_j \le t} \frac{\#\text{failures at time } t_j}{\#\text{at risk at time } t_j}$, $\hat{S}(t) = \exp(-\hat{H}(t))$. The former estimator is based on the multiplication law for conditional probabilities, while the latter combines hazard function estimates in a natural way. There is little difference between the two, except when the number of survivors is small. There are several variants of the Breslow estimator that take account of ties – the Efron estimator being one of the most popular.

There are several ways to incorporate explanatory variables into the survival or hazard function. In accelerated failure time models the distribution of the failure time, T, is multiplied by an acceleration factor dependent on the explanatory variables. The model is $S(t \mid \mathbf{X}) = P(T\exp(\mathbf{X}'\boldsymbol{\beta}) > t) = P(T > t\exp(-\mathbf{X}'\boldsymbol{\beta}))$. The proportional hazards model Cox (1972) works directly with the hazard function. The model is $h(t \mid \mathbf{X}(t)) = h_0(t)\exp(\mathbf{X}(t)'\boldsymbol{\beta}))$ where $\mathbf{X}(t)$ is a vector of explanatory variables, possibly time varying, and $\boldsymbol{\beta}$ is a constant coefficient vector.

The latter model is usually taken to be a semi-parametric model, and the baseline hazard function, $h_0(t)$, is not modelled. Cox (1975) suggested the method of maximum partial likelihood for estimating the parameters of this model. This method is to maximise the part of the likelihood function that contains the parameters.

Counting Processes

The theory of counting processes has been found useful for analysing survival models. Each transition for an individual adds 1 to the transition count. Sometimes, more than one transition is possible in one at-risk period, e.g. change of occupation, while for transitions from benefit to work the count is always 0 or 1, as the individual ceases to be at risk after a transition.

Martingale theory is a branch of counting process theory that is particularly useful in survival analysis. Apart from some technical conditions, a martingale is a process for which, given the current state, the expected change is zero. Sub- and super-martingales are similar, but the expected change for a sub-martingale is an increase, while for a super-martingale it is a decrease. (The term 'martingale' comes from the martingale system in gambling: increase the bet after each loss so that a win would recover all past losses. This, in turn, is derived from part of the harness of a horse designed to hold its head down to prevent it from bolting – via Spanish from the Arabic for 'fastening' and influenced by the town of Martigues in Provence, France.)

Residuals

The martingale residual is the difference between the observed count and the expected count under the model. If the model is correct then the martingale residuals form a martingale, and the sums of squared residuals, used in estimation of sampling variation, form a sub-martingale. Although the martingale residuals have a similar interpretation to the usual residuals in linear models, the analogy is not complete. For example, the residuals are not independent, although, relative to the true model, they are uncorrelated. This is not much help, as the true model is unknown. However, one use for martingale residuals is in fitting a proportional hazards model. A graph of the martingale residuals for a model with no parameters against an explanatory variable reveals the functional form of that variable (at least if the explanatory variables are uncorrelated). This works quite well if the correlations between explanatory variables are not too strong. This can also be used to discover how to modify a fitted model.

There are many kinds of residual, but we will need just one more: the Schoenfeld residual. These are useful for checking the proportional hazards assumption. A Schoenfeld residual is the difference between the value of an explanatory variable at a transition and its average value before the transition. The scaled Schoenfeld residual is defined to be the sum of the estimated coefficient and the residual divided by its variance. This should be independent of time if the proportional hazards assumption holds.

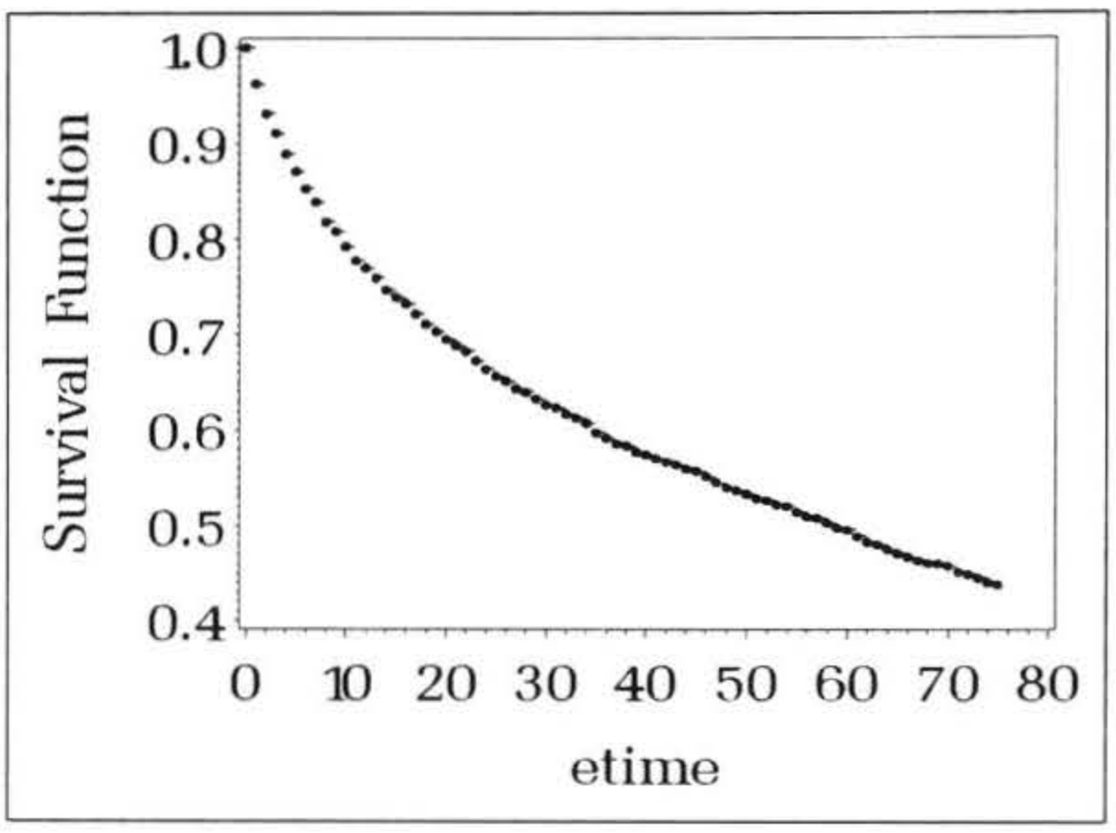
Exploratory Analysis

Figure 1 shows the Kaplan-Meier estimator of the survival function based on a model with no explanatory variables.

Note that the probability of remaining on a benefit declines steadily to about 0.5 after six years. In other words, of those who either remain on a benefit or move to

wages and salary (others being censored), about half will remain on a benefit after six years.

Figure 1: Kaplan-Meier Survival Function Estimate.



We next plotted the martingale residuals for this null model to assess the shapes of the functions representing the effects of the possible explanatory variables, e.g. age when first on a benefit. These functional forms are shown in Figures 2 to 6. The upper band of residuals corresponds to people who have moved to work while lower band corresponds to those who are still at risk.

Figure 2: Martingale Residuals for Sex.

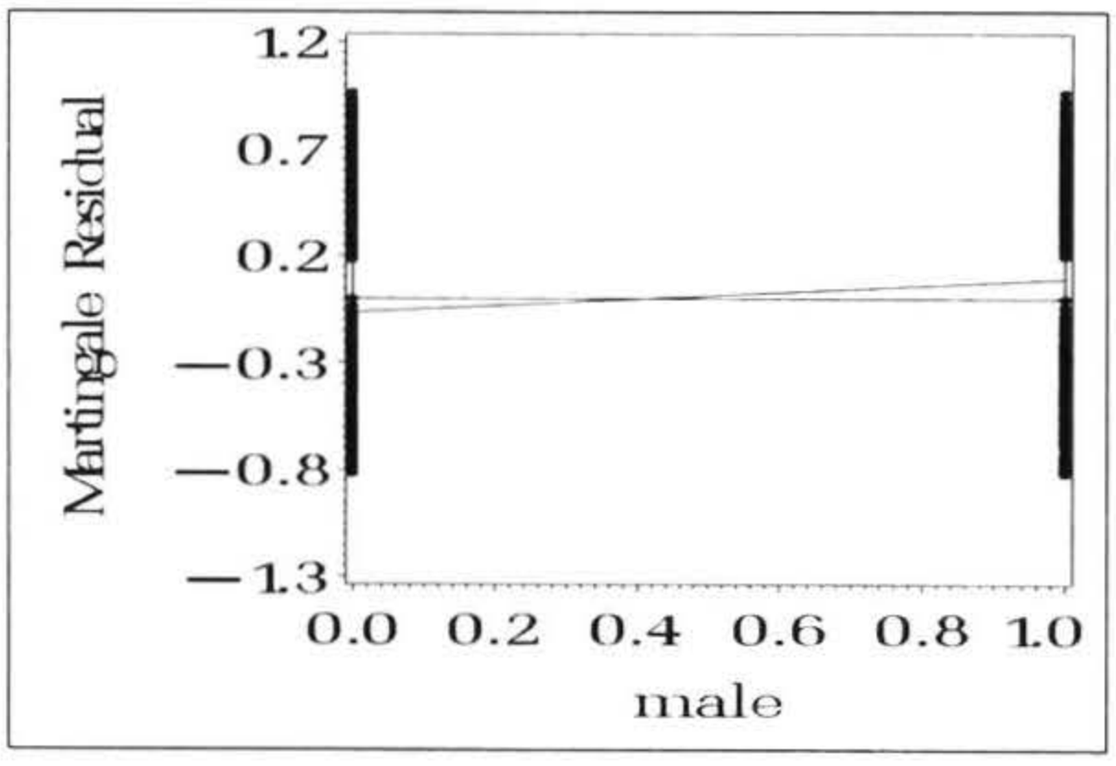


Figure 2 shows the effect of sex (female = 0, male = 1) – males being more likely to start work than females. Figure 3 shows the effect of age and sex – older people being less likely to start work. For age alone, the graph is very non-linear, but a quadratic seems to fit the centre portion. The graph flattens after age 58 years, while the part between age 18–30 years is a reversed S shape for which a cubic could be fitted. This flattening is likely to be due to the fact that many people receiving a benefit who are near the traditional retirement age stop seeking work. The different shapes of the curves for males and females mean that there is an interaction between age and sex. In particular, note that the female curve is very flat up to age 38. As the type of benefit is not available in the data, we may only speculate that many of the females are receiving a Domestic Purposes Benefit and are not available for work until their children are older. The flattening of the age graph at the end was due to the increased number of females, compared with males, over the age of 70 years who are still included in the data (perhaps partly due to greater life expectancy for females).

Figure 3: Martingale Residuals for Age.

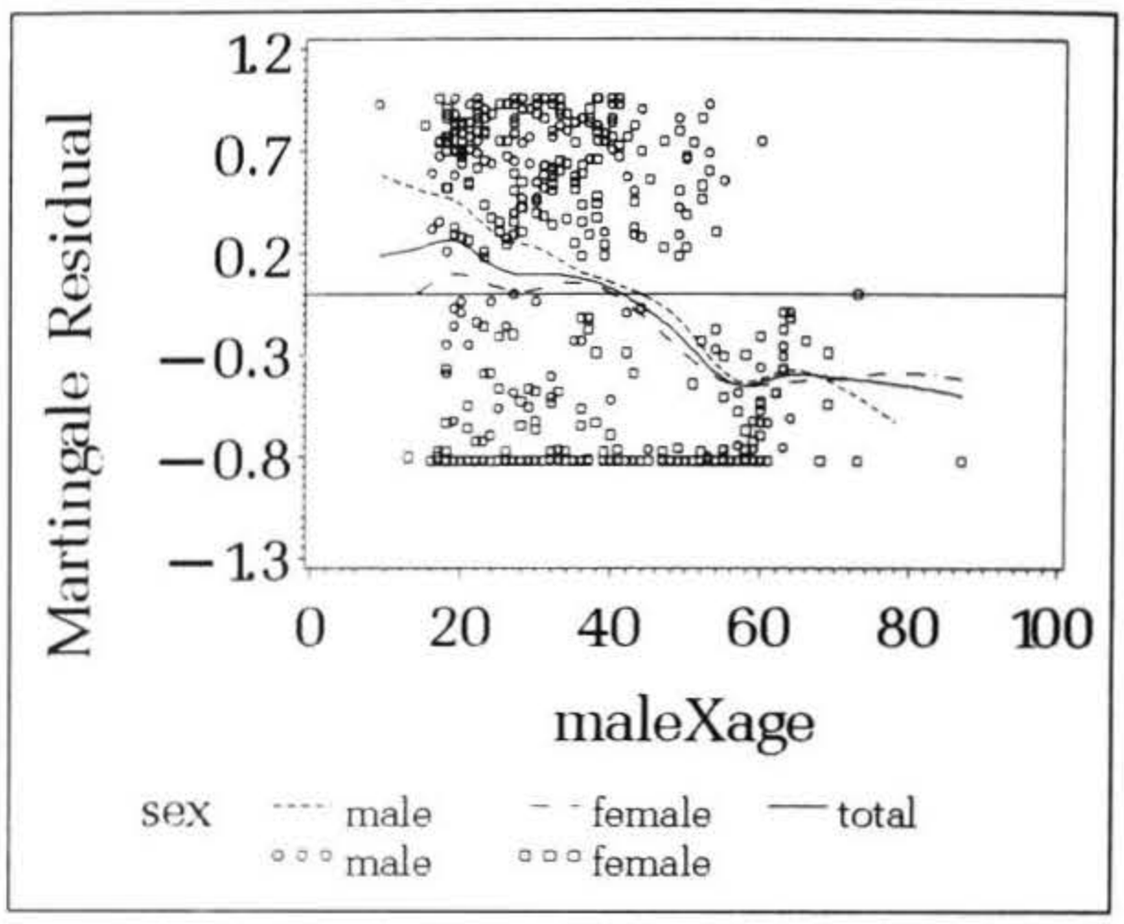
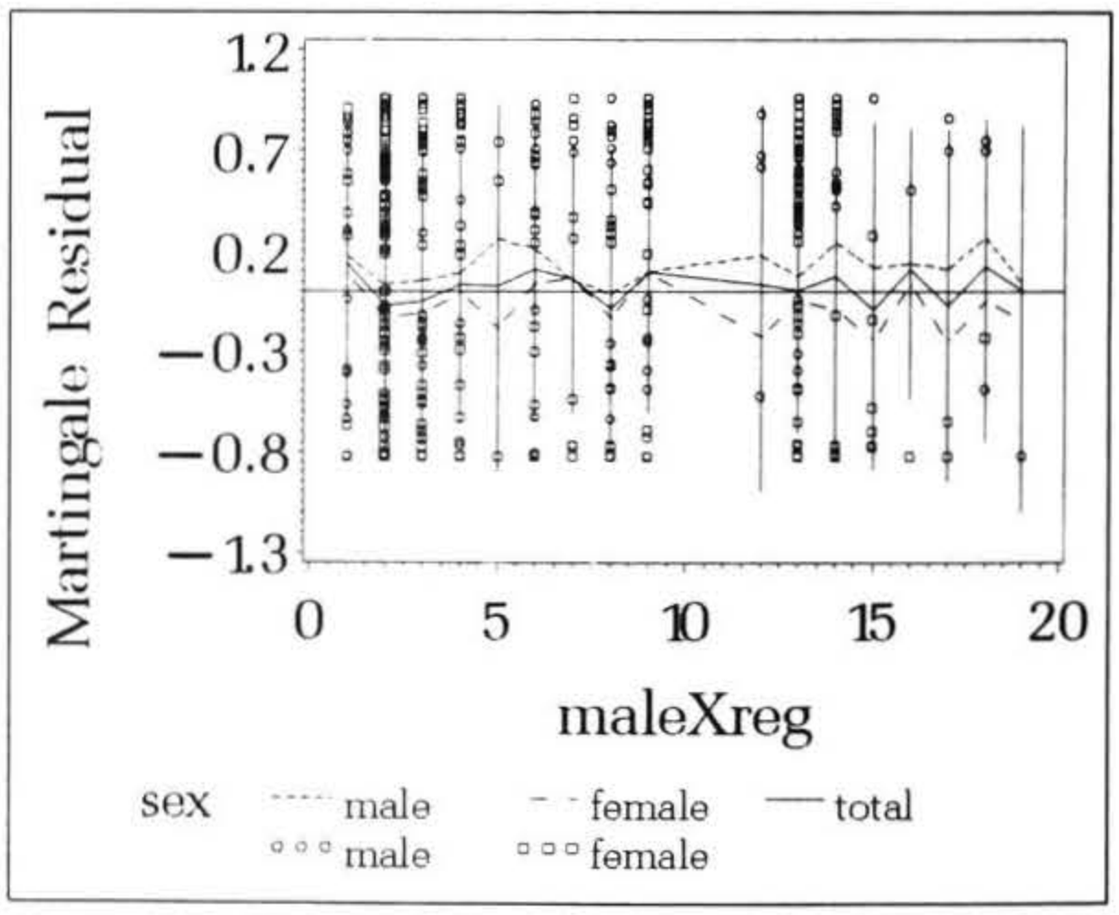


Figure 4: Martingale Residuals for Region.



The graph by region, Figure 4, shows that there are regional differences and that the shapes of the graphs are different implying that there is an interaction between the males and females. As the explanatory variable is categorical, the points have been joined with straight lines.

Figure 5 shows that there are differences by industry and that there is an interaction with sex. The small number of females in some industries, and males in others, accounts for much of the variation between sexes. The large numbers of females in retail trade and in accommodation, cafés, and restaurants is particularly notable. Those females have no previous industry shown in the dataset – either they have had no previous work or none during the time for which the data is available.

Figure 5: Martingale Residuals for Industry.

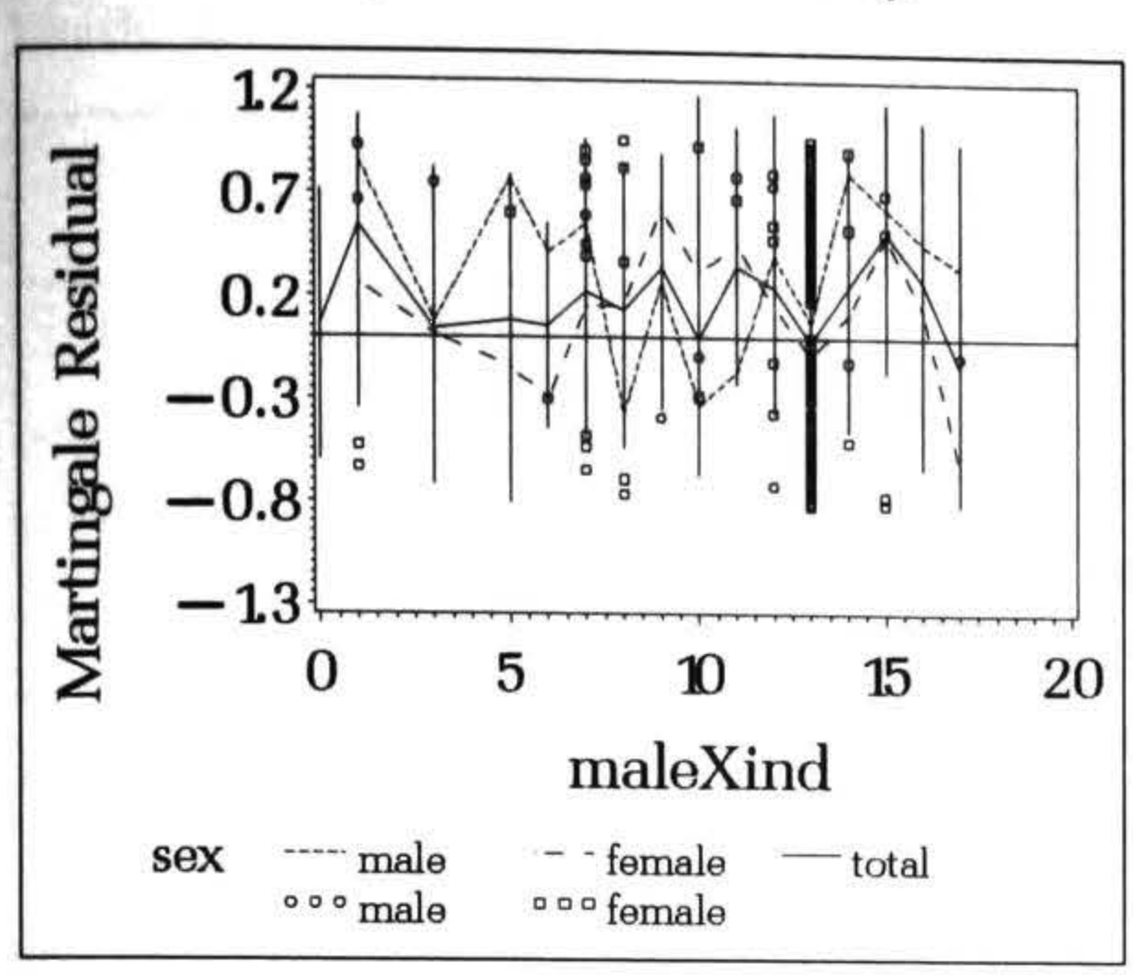


Figure 6: Martingale Residuals for Month.

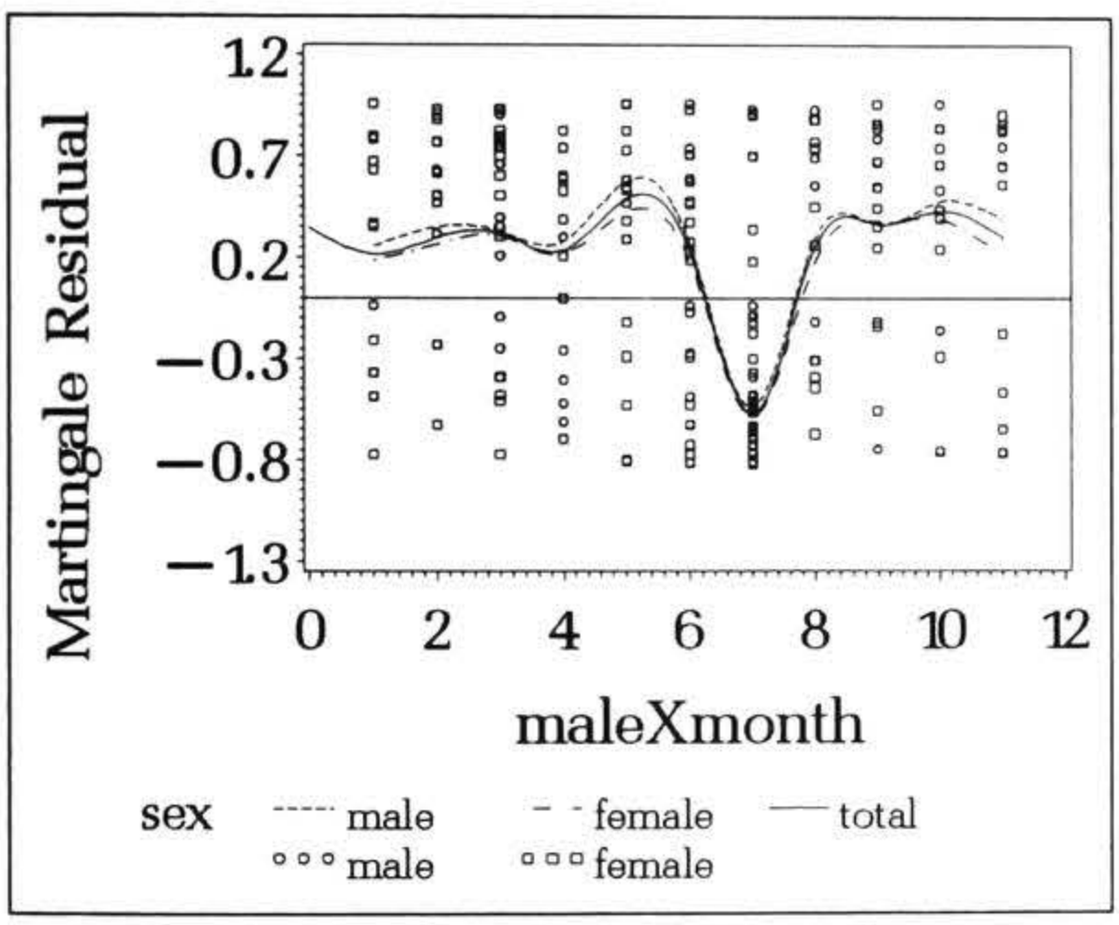


Figure 6 shows variation by month. Note particularly that month 7, which is August (January being 0), is very low. The shapes of the male and female part seem to be the same (with males slightly higher, as we found earlier). Thus, we chose a model with no sex by month interactions.

Fitting a Model

The preceding analysis suggests that we may use the following covariates: sex, male age, female age, male region, female region, male industry, female industry and quarter (grouping months into quarters). Age was coded as number of five-year age intervals greater than age 15 years. Using five-year intervals prevents the hazard ratio from being so close to 1 that it rounds to 1.000. We also added quadratic effects for male and female ages. These are coded as $\left(\frac{age-15}{5}\right)^n$ where n is 2 or 3. With this coding, the hazard ratio is the effect of a five-year change in age.

We used indicator variables to represent categorical data. A value 1 represents membership of the category and 0 represents non-membership.

The output from the analysis is shown below. Note that males have nearly two and a half times the probability of finding work than females (for any variable the baseline is the value 0). As noted previously, a very likely explanation is that many of the females are receiving a domestic purposes benefit and are not available for work. When a quadratic effect was introduced for female age the stepwise procedure did not include the male indicator variable. When this indicator was forced into the model the effect was reduced to 1.6. We also tried adding quadratic and cubic effects for both male and female age, but we only included linear terms in the final model. With a quarter effect in the model we obtained the results in Table 2. Note that the quarter effects are linearly dependent, so only three can be fitted.

Note that the hazard ratios change when new parameters are included. The reason is that some of the other variables are correlated. Therefore, adding variables to the model already partly accounts for their effect (the male effect increased because some of the correlations are negative). Note also that the fourth quarter effect is taken to be 1, which is the baseline. We can only estimate three-quarter effects unless we constrain their total effect – the constrained value becoming the baseline.

It is necessary to be careful with the interpretation of the hazard ratios for the continuous variable age. Consider a female aged 60 years. The hazard ratios for the variable age and those that we (rather inappropriately) called 'age squared' and 'age cubed' are 0.981, 0.813 and 0.922, respectively. The latter two are for differences in multiples of five years from age 40 years. Consider females aged 20 and 60 years. The value of age squared is 4 for both, while the values of age cubed are -8 and 8, respectively.

Therefore, the age components of the hazard ratios are, respectively,

$$0.981^{20} \times 0.813^{4} \times 0.922^{-8} = 0.573$$

and $0.981^{60} \times 0.813^{4} \times 0.922^{8} = 0.155$.

All other independent variables are indicators, so there is no such difficulty.

Table 2: Fitted Values from the Final Model.

The PHREG Procedure

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-square	Pr > ChiSq	Hazard Ratio
Male	1	-0.46823	0.47536	0.9702	0.3246	0.626
maleXage	1	-0.13958	0.02218	39.6071	<.0001	0.870
femaleXage	1	-0.22522	0.07825	8.2852	0.0040	0.798
qtr_1	1	-0.21973	0.08964	6.0083	0.0142	0.803
qtr_2	1	-0.10132	0.08393	1.4572	0.2274	0.904
qtr_3	1	2.63688	0.13483	382.4677	<.0001	13.970
qtr_4	0	0	.	.	.	.
femaleXreg1	1	0.50251	0.19580	6.5864	0.0103	1.653
femaleXreg8	1	-0.36758	0.17541	4.3915	0.0361	0.692
femaleXreg9	1	0.29580	0.12802	5.3385	0.0209	1.344
femaleXind7	1	0.57762	0.28519	4.1022	0.0428	1.782
femaleXind8	1	0.57429	0.24787	5.3681	0.0205	1.776
maleO57	1	-2.19723	0.27153	65.4824	<.0001	0.111
femaleO57	1	-3.60328	0.58536	37.8926	<.0001	0.027
femaleU37	1	-1.14097	0.46809	5.9413	0.0148	0.320
qtr_3Xtime	1	-0.08764	0.00328	714.7861	<.0001	0.916

Checking the Proportional Hazards Assumption

To check the proportional hazards assumption we consider a more general model in which the coefficients depend on time. The mean of scaled Schoenfeld residuals from the final model at a particular time is a measure of how far the time-varying coefficient differs from the one fitted by the proportional hazards model. A plot of this scaled residual against time for each parameter estimate should vary randomly about a horizontal line. Fitting a line to this data then reveals any time-varying nature for the parameter. There are too many graphs to show them all. A typical one is that for the variable male shown in Figure 7. It shows a more or less horizontal fit and therefore that the coefficient does not depend on time.

Some of the graphs show much more horizontal banding than this, but most do not reveal time-varying effects. However, a few graphs do reveal time-varying effects. For example quarter 2, the June quarter, as revealed by the graph in Figure 8. Evidently, the longer someone has been on a benefit and is still on a benefit in this quarter, the less likely they are to move to wages and salary compared with the December quarter base. This could mean that there is more chance of finding work in that quarter after a long period on benefit.

Figure 7: Scaled Schoenfeld Residuals for the Male Effect.

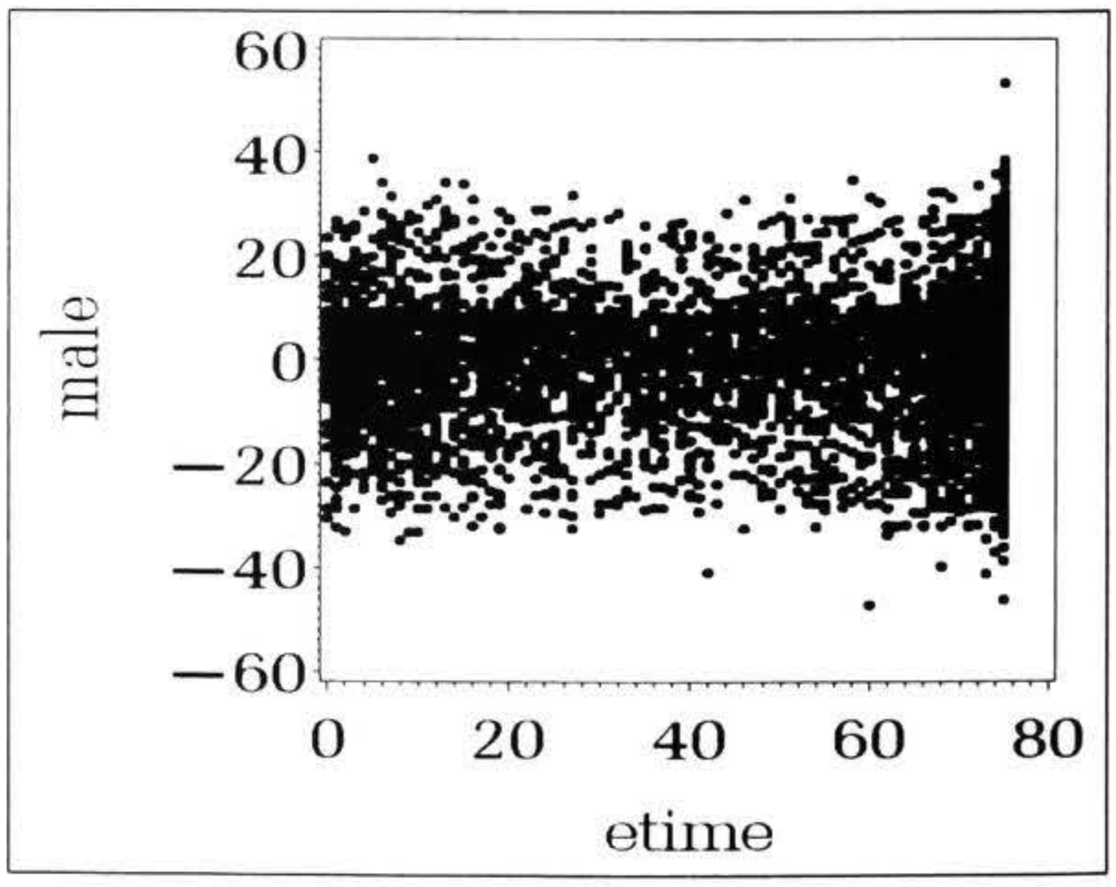
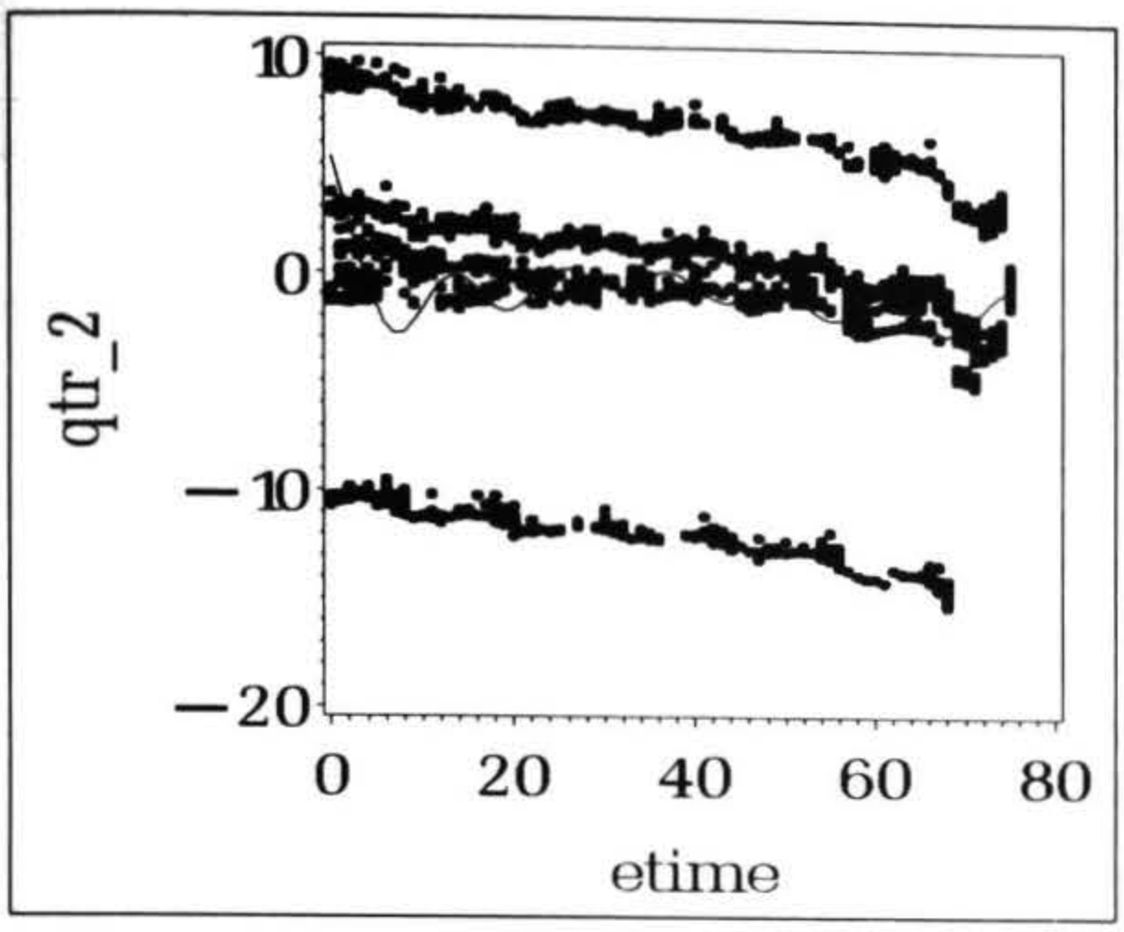


Figure 8: Scaled Schoenfeld Residuals for the Effect of Quarter 2.



Future Research

It might have been better to fit a curve to the effect of age rather than having threshold values at which there is a sudden change. Spline models (smooth curves made of pieces joined smoothly) could have been used. A cubic spline with just two knots could have been a better model.

There is a mixture of different types of individuals in the data: those whose first record was at age 15 years, those who had previously worked, and those who worked or were on a benefit before the data was collected. It could distort the results if these are kept in the same stratum.

We censored movements from benefit to a category other than wages and salary. This is a competing risks problem and could have been analysed as such.

There are people who have changed from benefit to work and back again more than once. These are correlated and this should be taken into account. However, there are only a few of these, so they should make little difference.

Apart from the known heterogeneity in the data, there are variables that we might have liked to have observed but which were not available. We might expect, for example, that ethnicity has an effect on transitions. Other variables that we have not even thought about might also have significant effects. It is possible to model such effects by assuming that each individual has an individual effect to be added to the other effects. Although this effect (known in the reliability literature as 'frailty') is unknown, we often assume a simple model for it and estimate the model parameters. A gamma distribution is often assumed. This adds two extra parameters to the model, allowing a little more flexibility in assessing the variation between individuals.

Notes

1 Any views expressed are those of the author and do not purport to represent those of Statistics New Zealand. Any remaining errors are the sole responsibility of the author.

2 The tables in this paper contain information about groups of people so that the confidentiality of individuals is protected. Only people authorised by the Statistics Act 1975 are allowed to see the data about a particular person or firm. The results are based in part on tax data supplied by the Inland Revenue Department (IRD) to Statistics New Zealand under the Tax Administration Act. This tax data must only be used for statistical purposes and no individual information is provided back to IRD for administrative or regulatory purposes. Any discussion of data limitations or weaknesses is in the context of using the Linked Employer-Employee Data (LEED) for statistical purposes, and is not related to the ability of the data to support IRD's core operational requirements. Careful consideration has been given to privacy, security and confidentiality issues associated with using tax data in this project. A full discussion can be found in the LEED Project Privacy Impact Assessment paper Statistics New Zealand (2003).

3 I thank Sarah Crichton, Walter Davis, Sylvia Dixon, Tas Papadopoulos, Steve Stillman and participants at Statistics New Zealand LEED Research Forum for discussions and valuable comments.

References

Allison, P.D. (1995). *Survival Analysis Using the SAS System: A Practical Guide*. SAS Institute.

Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, **34**, 187–220.

Cox, D.R. (1975). Partial likelihood. *Biometrika*, **62**, 269–276.

Beamonte, E. and Bermúdez, J.D. (2003). A bayesian semiparametric analysis for additive hazard models with censored observations, *Test*, **12**(2), 347–363.

Carroll, N. (2006). Explaining unemployment duration in Australia, *Economic Record*, **82**(258), 298-314.

Gallo W.T., Teng, H.M., Falba, T.A., Kasl, S.V., Krumholz, H.M. and Bradley, E.H. (2006). The impact of late career job loss on myocardial infarction and stroke: a 10 year follow up using the health and retirement survey, *Occupational and Environmental Medicine*, **63**, 683-687.

Hyslop, D., Stillman, S. and Crichton, S. (2004). *The Impact of Employment Experiences and Benefit – Spell Duration on Benefit-to-Work Transitions*. Statistics New Zealand, LEED Project Technical *Paper*.

Knut, R. and Zhang, T. (2003). Does unemployment compensation affect unemployment duration? *Economic Journal*, **113**(Jan), 190–206.

Lüdemann E., Wilke R.A. and Zhang, X., (2005). *Censored Quantile Regressions and the Length of Unemployment Periods in West Germany*. Discussion Paper No. 04-57.

Moore, T. (2004). Longitudinal analysis of labour force data. *Labour Employment and Work in New Zealand, Proceedings of the 11th Conference*. 205-210.

Therneau, T.M. and Grambsch, P.M. (2000). *Modelling Survival Data: Extending the Cox Model*. Statistics for Biology and Health, Springer.

Statistics New Zealand (2003). Linked Employer-Employee Data Project: Privacy Impact Assessment, http://www.stats.govt.nz/NR/rdonlyres/F5025B36-85D8-4464-B683 CE070FFC4807/0/LEED PrivacyImpactAssessment.pdf

Author

Terry Moore
Statistical Analyst
Statistical Methods
Statistics New Zealand
P.O. Box 2922
Wellington
Terry.Moore@stats.govt.nz