LONGITUDINAL ANALYSIS OF LABOUR FORCE DATA

Terry Moore

Statistics New Zealand

Abstract

The Household Labour Force Survey (HLFS) is a quarterly panel survey that is used to provide a snapshot of the New Zealand labour force at a point in time. Although originally intended for cross-sectional purposes, the fact that the occupants of the same households are interviewed for up to eight quarters makes it possible to extract longitudinal information, such as labour force dynamics. This paper will discuss some of the longitudinal uses of the data, and some potential problems and issues.

One issue discussed is data cleaning. The HLFS imputes some variables, such as age and sex, when data has not been provided by respondents. However, as the main objective is to produce cross-sectional estimates, there is no attempt to achieve longitudinal consistency, so apparently people may change sex or be rejuvenated. I will discuss some methods for cleaning the data and investigate whether this has any significant effects on longitudinal estimates, such as hazard rates.

The results suggest that it is feasible to obtain longitudinal information about transitions in labour force status from the HLFS data, but it is necessary to concatenate panels to obtain reasonable accuracy. Editing seemed to make little difference to the conclusions.

Introduction

The Household Labour Force Survey (HLFS) is a quarterly panel survey with about 15000 households being surveyed each quarter. An attempt is made to interview all the individuals in each selected household each quarter for the life of the panel. The cross-sections all consist of eight panels. Except when a new sampling frame is being introduced, the panels remain in the study for two years and are rotated in and out at quarterly intervals. This is also true of the latest redesign, but previously some panels have been shortened. Hopefully this will be avoided in future.

The main focus of the HLFS is the labour force status (employed, unemployed, not in the labour force, and three categories for not valid) and variables that are likely to explain transitions between these states. Covariates, or explanatory variables, are age, sex, marital status, ethnicity, education etc., as well as the past history of labour force status. Whereas a cross-sectional study can only consider the state of all the variables at a particular time as an explanation of the current status, a longitudinal study can explore the whole past history for each individual. Both cross-sectional and longitudinal analyses allow us to relate numbers or proportions of people in the various states to their explanatory variables in each quarter, as well as to consider the changes between quarters. However, the changes in a longitudinal analysis are *gross flows* (amount of movement between states) rather than the *net flows* (total change in a state—the difference between movements in and out of the state) that a cross-sectional analysis provides. In other words, if the numbers of people moving between states balance, leaving the number in each state approximately constant, there could be considerable change in individual employment status that a cross-sectional analysis will not detect. Furthermore, any time varying explanatory variables are less informative in a cross-sectional analysis because their past history cannot be used.

If we know all the gross flows, we can easily calculate net flows. Simply add all the gross flows into each state and subtract the gross flows from the state. After estimating gross flows and net flows we may check the longitudinal models used by comparing the resulting net flows.

We considered two complementary types of analysis: gross flows between states, and hazard rates (also known as failure rates) for transitions between states. Whereas net and gross flows are related concepts, hazard rates, which belong to the realm of survival analysis, have no cross-sectional counterpart.

The following example illustrates the difference between net and gross flows. If the employment (E), unemployment (U) and not in the labour force (N) rates are 70%, 10% and 20% respectively, and subsequently become 75%, 7%and 18% then net flows to these categories are +5%, -3% and -2%. This could represent gross flows of 3% from U to E, and 2% from N to E. But it could equally well represent 5% from U to E, and 2% from N to U, or, among other possibilities, 10% from U to E, 20% from N to E, 7% from E to U, 18% from E to N and 0% in either direction between N and U.

Terminology

The traditional vocabulary of survival analysis evolved from studies of failures of some kind, such as deaths, injuries, industrial action or vehicle breakdowns. It seems rather negative when applied to positive events such as becoming employed but there is no obvious positive or neutral vocabulary. When in some state (unemployed, for

example) we say that an individual is *at risk* of making a transition to another state. The set of individuals at risk over any given period (or instant of time) is called the *risk set* for that period. For an at risk individual, the time up to a transition is the *survival time* in the state, and the probability that an individual will still be in the state at a particular time is the *survival probability* for that state and time. The *hazard rate (failure rate)* for a transition is the conditional probability density of the time to a transition given survival to that time. I.e. the hazard rate, $h(t)$, is given by $h(t) = f(t)/S(t)$ where $f(t)$ is the probability density for the transition at time t and $S(t)$ is the *survivor function* (probability of survival to time t or greater). The *hazard function* (hazard rate as a function of time) is not a probability density function because for each value of t we condition on a different event—survival up to that time. The integral of the hazard rate from time zero, as a function of the upper limit, is known as the *cumulative hazard (failure) rate* and is increasing and generally unbounded. For a continuous distribution, if $F(t)$ is the cumulative distribution function for transitions, and $H(t)$ is the cumulative hazard rate, then clearly $S(t) = 1 - F(t)$, $H(t) = -\ln S(t)$ and $f(t) = h(t)\exp(-H(t))$. Thus if any of $f(t)$, $F(t)$, $S(t)$, $h(t)$ or $H(t)$ are known up to some given time, then all can be found up to that time. (In the discrete case, the hazard rate is the probability of failure at an instant given survival up to that instant. But the formulae are not quite as simple as in the continuous case.)

We are likely to be interested in a particular transition (to employed, for example). Given that there is a transition, each new state has a certain probability. We require the probability density for the transition of interest to us. This is the product of the probability density for a transition and the conditional probability of the transition of interest given that there is a transition. We can define the hazard rate for this specific transition to be the density for this transition divided by the survivor function. See the discussion of deficient distributions, however.

The shape of the graph of the hazard function is very informative. If the hazard rate for becoming employed is increasing then the longer one has been unemployed, the more hopeful the situation becomes. The reverse is true for a decreasing hazard rate. The neutral case (constant hazard rate) is a convenient standard for comparison. In this case transitions are completely random and a transition to employment is pure luck. This is the well known Poisson process that represents a sequence of independent random events, the number of events in a given interval having the Poisson distribution and the time from any origin to the next event having an exponential distribution.

When there are only two states, the at risk individuals must make a transition to the other state or remain at risk. In our labour force study, there are five states. Therefore the individual could cease to be at risk of a particular transition without making the one that interests us. If, for example, we are interested in transitions from unemployed to employed, there could be a transition to not in the labour force. This is an example of *right censoring*. Observations could be censored for other reasons such as non-response, leaving the surveyed household, or their panel being rotated out.

Data may also be *left censored* because the employment status is unknown before the start of the study and some individuals move into a surveyed household during the life of the panel. Unfortunately, censoring can cause problems with the analysis. The issue is whether censoring is *non-ignorable* or *ignorable*, in other words, whether or not the distribution of a variable, given the available observations, depends on unobserved variables. We do not know the employment status of people who moved away. Ignoring them when they moved to a job would downwardly bias our estimates of flows to employment. However, ignoring them when they moved for some other reason would cause an upward bias. It is reasonable to assume that censoring at the beginning or end of the panel is ignorable. Otherwise it is likely to exert some influence on the results, but we cannot say how much. Fortunately, in the HLFS, unemployed individuals are asked how long they have been seeking work, and how long they have been unemployed. This ameliorates the problem with left censoring.

Deficient Distributions

If there are individuals who, after some time, have absolutely no chance of making a transition, then the total probability could be less than 1. We could interpret this to mean that a transition might never occur, or we could think of it as occurring after an infinite time. The distribution is said to be *deficient* if the survivor function tends to a positive limit as t increases. As our subjects are only studied for eight quarters (and, in any case, no-one is immortal), we cannot necessarily distinguish between a deficient distribution and one with long survival times. However, deficient distributions will occur when there are more than two states because a transition from state U to E (for example) precludes a transition from state U to N. If the latter transition has a positive probability of occurring sometime, then the former transition has probability less than 1 of occurring at all. Of course it might be possible to enter the state U again, but we are referring to this particular risk set, not the next. However, if N stands for 'not in the labour force' then there are some people for whom that will be an absorbing state (for example those who retire). Therefore, even taking multiple at risk periods into account, we expect some degree of deficiency in the distributions.

Group versus Individual Hazard

Some people have more chance of a transition and therefore a greater hazard rate than others. This is true even for those with identical values of the explanatory variables except in the unlikely event that we have included all variables in our model that influence this probability. We cannot estimate the individual hazard rates, all we can do is to give a sort of average, or

marginal hazard rate, for the group. There is a danger of misinterpreting this marginal hazard. For example, a hazard rate is often observed to be 'bathtub' shaped—it decreases and then increases. If we are talking of breakdowns of vehicles, the increase is easy to explain in terms of wearing out. However, we must guard against misinterpreting the initial decreasing hazard rate. It is doubtful that individual vehicles become noticeably more reliable as they are 'run in' (although there could be some slight effects, e.g. work hardening or 'bedding in'). The truth is more likely that some vehicles already have faults when they leave the factory. Some of these vehicles are more likely to fail early, and this leaves a population of vehicles that are more robust on average. In the case of labour force status transitions, a decrease in the hazard rate for becoming employed need not mean that individuals tend to give up looking for work. No matter how plausible that hypothesis might be, those who are more likely to find employment are removed from the risk set. Unless the hazard rate is homogeneous between individuals, this must account for at least some of the decrease. Unfortunately, heterogeneity of hazard rates between individuals is difficult to detect unless a large number of individuals have repeated events. Even then, in order to detect the heterogeneity we must make assumptions about the repeated event process.

Limitations of Quarterly Data

As data for each household is collected during the same week, it is possible that we might overestimate the length of a spell if there were two or more changes of state between interviews. There are two scenarios in which this will cause a large bias—if there is a great deal of churning (frequent changes between states), or if many spell lengths are less than one quarter. In the latter case we will fail to observe those who become unemployed and employed again between interviews. Although we can estimate how many such people are missed, the shorter the spell the greater will be the error in this estimate.

The granularity limitation can be overcome by asking the subjects to recall information from between interviews as is done in the Survey of Family Income and Employment (SoFIE). But this introduces another problem: how well can people recall the information?

Weights

As the HLFS data is obtained form a complex survey design (multi-stage cluster), it is usual to take account of the selection probabilities using weights. Initially the weights are the reciprocals of the selection probabilities. Further adjustment is made for non-responses (post stratification), then the members of each household are given equal weights using a regression method (integrated weighting). As the weights were designed to reduce bias for cross-sectional estimates of means, it is not clear what weights to use for longitudinal estimation. Also, there is some debate in the literature about whether selection weights should be used in regression modelling. Some studies have suggested that it makes little difference. It is also suggested that fitting a model to the covariates makes the weights redundant. Suppose that we stratify by income. Then we sample more individuals in some income groups than others. A scatter plot will reveal clusters of incomes. The regression will automatically be weighted towards the larger clusters, as it should be because we have more precise information in those regions. It is valid to use weights to take account of the variance of the response variable about the regression and, especially when repeated measures are made on the same individuals, covariances between them. But these are not the selection weights. If we were not using a regression, we would also need to take account of the number of individuals in the population in each stratum, but in a model-based approach we are taking account of the explanatory variables, and this indirectly takes account of the strata.

Little (2004) says: "Survey sampling is perhaps unique in being the only area of statistical activity where inferences are based primarily on the randomization distribution rather than on statistical models for the survey outcomes." Regarding the difference between model-based weighting and design-based weighting, he continues: "Both forms of weighting seem plausible, but they are not necessarily the same."

The method of weighted least squares gives the best linear unbiased estimator for a linear model when the weights are known and can be expected to perform well when the weights are estimated from the data. But these weights are not the selections weights. They are based on the distribution of the residuals from the regression line.

Little (2004) gives a compromise between the two weighting schemes but we have made no attempt in this direction.

Data Cleaning

Missing values of a few variables, sex and age, for example, had already been imputed by the time we received the data. As the survey was originally designed as a cross-sectional study, there was no attempt to make the data longitudinally consistent. Additionally, there are some errors in these variables as well as in the non-imputed variables. In some, but not all, cases it is possible to correct these values with little likelihood of error. For example, we can be almost sure of how to correct an age that annually increases by one except for one stray observation. On the other hand, if the age seems to change at random, it is unlikely that we can correct it reliably. It could be helpful to know how much difference data cleaning makes to our analysis. If there are only a small number of errors, ignoring them might make little difference to our conclusions. Even if some of the explanatory variables have impossible longitudinal changes, a model can still be fitted (probably with more outliers).

For panels from quarter 45 (December, 1996), but not for earlier panels, imputed variables have imputation flags. For cross-sectional work, missing values of sex had been imputed randomly. For this analysis I took the most common value of the sex for each individual and imputed randomly in the event of a tie. There were very few such cases. Missing ages had been imputed for everyone not flagged as a child by first matching a five year age group with a similar subject, then choosing an age randomly in the group. I fitted a robust line of gradient 1/4. More precisely, I used $Q/4 + A$ where A is the median of $age - Q/4$ and Q is the number of the quarter.

Multiple ethnicities are allowed. The HLFS records up to three values for each quarter. As ethnicity is based purely on the subject's perception, it need not be constant. However, it seems unlikely that those who change their perceived ethnicity would be very different from those who do not. And given that some errors or imputations would have occurred, I decided to make the ethnicity constant. Some recording errors were obvious. When a new category (other European) was added, some field staff seemed to use the old coding forms for varying numbers of quarters. This was obvious from the pattern of responses adjusted to the new codes. Another obvious change was for those who switched to the new category. Unfortunately, it is not possible to infer who would have used that category when it was not available. After fixing the obvious errors, of all the ethnicities given by each subject over eight quarters, the three most common were chosen.

Some changes in marital status are impossible, for example from married to never married. Married is interpreted as *living* as married. If never married is interpreted as never *formally* married, there is scope for different interpretations. How would separated, divorced or widowed be interpreted by someone who has ceased living as married but has never been formally married? Nevertheless I removed those transitions I deemed impossible or unlikely by using the likely response from the closest quarter.

The variables 'weeks seeking work' and 'weeks without work?' are difficult to clean. After a spell of employment they must drop to less than 13 weeks and then rise by 13 each quarter. But for those unemployed at two successive interviews who give a number of weeks without work as less than 13, we could assume that there was a spell of employment between interviews. However, if the number of weeks jumps by more than 13 at the next quarter, one of the responses must be incorrect. A reasonable method would be to use a piecewise linear function, with each piece having gradient 13 and dropping to less than 13 at each break point. Determining the break points is difficult but might be possible using dynamic programming so long as we can choose a suitable measure of goodness of fit.

Similar methods can be used for other variables, but for this study we restricted ourselves to these.

Imputation

In most cases the adjustments discussed in the previous section are likely to be reliable because the values can be checked against data for other quarters. There is also a large amount of missing data for which imputation is less reliable. In previous studies of the HLFS data set, these values were imputed using a non-parametric modelling method. The gist of the approach is to identify similar individuals based on any observed variables, and then to choose values at random from the records of these similar individuals. In some previous studies the tree based subset selection method, CHAID was used. Imputation is particularly attractive when only a few items are missing from individual records, but a large number of records have missing values. Ignoring the records with missing values wastes much valuable data, while imputation only manufactures a small amount of 'information'. However, when a large amount of data is missing, imputation is less attractive as it tends to bias results towards those obtainable from the complete cases. But ignoring records with missing values also causes bias, but in the opposite direction. The issue is whether non-response is *ignorable* or *non-ignorable*. In the former case, the population who fail to respond are very different in the characteristics that interest us from those who respond.

Imputation has another problem: if we impute randomly we could reduce the effect of any correlation between variables for the same subject, while if we use the mean values for the similar individuals we tend to increase that correlation as well as making the variance appear less than it really is.

Kuzmicich and Wigbout (2001) solved this problem by imputing missing values from similar individuals who were matched using the tree based method CHAID.

Types of Non-response

In item non-response, some variables are missing, in unit non-response, no data exists for a person (or too little is available to impute the rest). For longitudinal surveys, a third type of non-response is wave non-response, in which no data (or not enough data) exists for a person for a particular time period. For item or wave non-response we can use other time periods for imputation using the same methods as for correcting inconsistencies.

Inference

Even though the HLFS data sets are large, many of them having between 2000 and 2500 dwellings and between 8500 and 9000 individuals, many of the individuals are children or not in the labour force for some other reason. On further restricting the set to those who are unemployed at some time during the life of the panel, the sample size reduces to between 700 and 800. The HLFS has a large number of variables, but we cannot fit a model to very many of them without much more data.

We wish to estimate the effects of certain variables on the hazard rate for the transition from unemployed to

employed. The covariates used were marital status, age, sex, ethnicity, has school qualification, has post school qualification, and local government region. Additionally, there might well be an effect due to the changes in employment rates over time. Thus we should also allow a time effect.

The original plan was to use each panel separately to estimate the effects of the covariates, and then to smooth the coefficients in the model over time. However, within a single panel of data the variables were not significant. In previous studies, panels had been combined, but this tends to blur any time effects, especially seasonal effects. And seasonal effects are potentially important—someone still unemployed at a season of low unemployment might expect a greater chance of becoming employed than at a time of high unemployment. To overcome this problem we added three more variables, the effects of three of the quarters. With four observations per year we cannot estimate more than three periodic terms. The revised plan was to combine eight panels at a time using a sliding window and smooth the coefficients of the model, including the periodic variables. This allows us to examine time trends in the variables as well as seasonal effects. As up to seven panels overlap in the different windows, the estimates of the coefficients are not independent. This does not nullify the estimates, but it does do *a priori* smoothing before applying time series methods. This also has an effect on the apparent variability of the estimates.

Hazard Models

The two most commonly used models in survival analysis are the *accelerated failure time model* and the *Cox model*. In both models the covariates determine the degree of departure for an individual's hazard rate or survival function from the average for the data as a whole. The average is known as the *baseline*. In the accelerated failure time model the failure time is multiplied by a factor dependent on the covariates. In the Cox model, the hazard rate is multiplied by such a factor. When the covariates are not time dependent, this is known as the proportional hazards model, and the ratio of hazards for any two individuals just depends on the covariates and not on the baseline hazard function. We fitted the Cox model using SAS which allows for time dependent covariates and multiple at risk periods (an individual can be unemployed more than once). With this model SAS does not produce a baseline hazard function. However, the main purpose is to see what effect the covariates have on the hazard rate. The precise form of the Cox model is $h(t) = h_0(t) \times \exp(\sum x_i(t)\beta_i)$ so that the logarithm of the baseline hazard rate is increased by a linear function of covariates. The *hazard ratio* is the factor by which the hazard rate is multiplied if a covariate is increased by 1. For example, the variable *sex* is an indicator variable that is 1 for a female and 2 for a male. If the hazard ratio for this variable is 1.05, for example, then males would be 5% more likely to have a transition than females, other variables being equal.

Results

Some of the covariates do seem to have an impact on the hazard rates for a transition from employed to unemployed. However the effect might not be significant in every sliding window. In order to smooth a series of effects over time it is necessary to use an estimate whether or not its quality is good. When an effect is not significant, its standard error will be large. Because adjacent windows have seven quarters in common, they are expected to be similar Smoothing should, however, improve the quality.

Variables that have a significant positive effect for some windows include male, married and European. Those that sometimes had a significant negative effect include age, school qualification, post school qualification, Pacific ethnicity, and living North of Auckland. Seasonal effects, when significant, were not in consistent directions. A sample of the SAS output for March 1989 is shown in table 1. Only two of the variables showed a steadily changing trend, sex and post school qualification. The graph in figure 1 shows an increasing trend in the effect of school qualification from negative to slightly positive. The positive effects in the recent part of the graph were not significant, however. Because of the coding used for the qualifications variable, negative means that it is an advantage to have a qualification. The effect of sex seemed to decrease from 1990 so that the advantage for males diminished. The effect of age was at a minimum around 1990 to 1997.

Table 1: Output for March Quarter 1989

The SAS System 08:34 Thursday, November 18, 2004 231

The PHREG Procedure

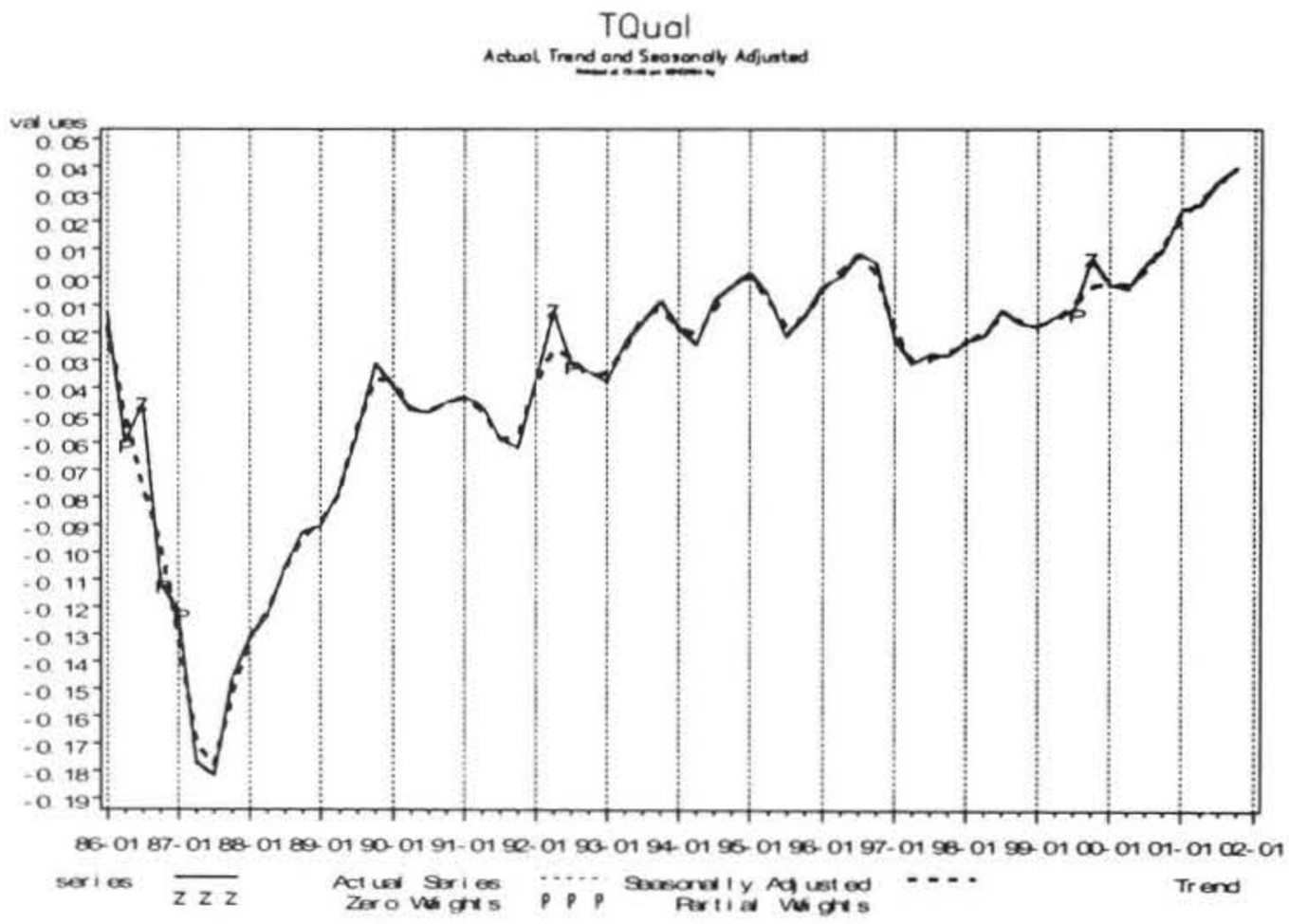
Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	158.8665	15	<.0001
Score	157.5464	15	<.0001
Wald	156.5215	15	<.0001

Analysis of Maximum Likelihood Estimates

Variable	Parameter DF	Standard Estimate	Error	Chi-Square	Pr > ChiSq	Hazard Ratio
sex	1	0.17004	0.03789	20.1364	<.0001	1.185
age	1	-0.00628	0.00214	8.6366	0.0033	0.994
msi1	1	0.05650	0.06793	0.6918	0.4056	1.058
msi5	1	-0.04429	0.07916	0.3130	0.5758	0.957
qMar	1	-0.09943	0.05828	2.9105	0.0880	0.905
qJun	1	-0.06914	0.06070	1.2975	0.2547	0.933
qSep	1	0.15428	0.05939	6.7494	0.0094	1.167
SQual	1	-0.11432	0.04074	7.8762	0.0050	0.892
PSQual	1	-0.09095	0.02047	19.7421	<.0001	0.913
ethi1	1	0.03228	0.12957	0.0621	0.8033	1.033
ethi2	1	-0.18016	0.13728	1.7223	0.1894	0.835
ethi3	1	-0.31159	0.14275	4.7647	0.0290	0.732
LGRi1	1	-0.08614	0.05256	2.6862	0.1012	0.917
LGRi2	1	-0.00969	0.05912	0.0269	0.8698	0.990
LGRi3	1	-0.02868	0.05760	0.2480	0.6185	0.972

Figure 1: Trend for Effect of Post School Qualification



Conclusions

The HLFS data set provides a rich set of data for longitudinal analysis. However, it can be rather limiting for some types of analysis because of the small number of people in some subgroups. However some conclusions could be drawn. Data cleaning seems to make little difference.

Further Research

Further research could try to identify which transitions provide the greatest amount of useful information. It would also be useful to find methods for fitting a baseline hazard rate when there are time varying covariates.

Transitions from employed to unemployed could also be worth investigating. Other possible covariates could be included, such as characteristics of other household members.

References

Little, R.J. (2004). To Model or Not To Model? Competing Modes of Inference for Finite Population Sampling. *Journal of the American Statistical Association*, **99**, 466, 546-556.

Kuzmicich, G. and Wigbout, M. (2001). A longitudinal look at some data of the Household Labour Force Survey. *Research and Analytical Report* **16**, Statistics New Zealand.