Effects of Measurement Error on Occupational Event History Analysis

Daniel H. Hill[1]
Survey Research Institute, University of Toledo, Toledo, OH 43606

Introduction

Event history analysis is one of the more promising new tools available to academic and policy analysts interested in the dynamic processes underlying employment and program participation behavior. The monthly dating of occupation and industry employment, along with its large size and national representativeness, potentially make the SIPP one of the most attractive data bases for such analyses. The extent to which the well documented measurement errors in the SIPP (and all other panel studies with reference periods longer than the basic unit of measurement) detract from this potential is not well understood. The natural experiment resulting from the overlap in the 1985 and 1986 SIPP Panels, which differ only in the **method** of collecting occupation and industry data, provides a unique opportunity to gain insight into the effect of measurement errors on event history analysis.

Earlier analysis (Hill, 1992) found that the amount of month-to-month gross change in the SIPP occupation data is roughly six times as great when the questions are asked and coded each wave (independent collection method), than when they are only asked if the respondent reports a change in duties (dependent method).

Conventional wisdom suggests that the dependent method is inferior because it results in many more "false negative" measures of change than does the independent method. For occupation and industry change, however, there is some evidence that dependent methods actually improve the signal-to-noise ratio of the data because the reduction in "false positive" measures more than offsets the increased false negatives. Evidence of this is particularly strong for industry change. With independent data collection, the SIPP data indicate nearly 51 million month-to-month industry changes during the 2/86 through 4/87 period for individuals working for the **same employer** in both months. With dependent collection, this figure is reduced to less than 5 million changes. The number of industry changes for individuals who also changed employers is 18 and 17 million during the period for independent and dependent modes, respectively.

In the present paper we will attempt to assess the relative empirical validity of the occupation change measures for the two methods in the context of event-history analyses. To do so we need external knowledge. Lacking direct observations of "true" occupation change, we are forced to use other forms of knowledge to serve as benchmarks in assessing the relative validity of the measures.

Our confidence in these external benchmarks varies considerably. The best source of external knowledge available to us is a consequence of the basic sample design. Specifically, the SIPP has been designed so that independent quarter samples are interviewed monthly on a rotating basis. This means that no matter what the calendar-month pattern of true change, in the absence of measurement error, each month of the four-month reference period should contain roughly one-quarter of the observed change in occupation. The extent to which change clusters at the seams of consecutive reference periods, therefore, provides a very sound method of assessing the relative quality of the data from the two collection methods.

The second source of external knowledge upon which we can assess validity is the extent to which observed change is associated with either other observed changes or with the characteristics of sample members and/or their occupations. True change in occupation "should" be related to change in wages and "should" be more prevalent for younger individuals whereas false change need not be. The extent to which one data collection method produces stronger associations (in the "right" directions) with these measures than the other method is, therefore, a measure of its relative empirical validity. The accuracy of the assessed relative validity depends, in this case, on the accuracy of our a priori knowledge. Nevertheless, if the differences in the results obtained under the two collection modes are large, we should be able to distinguish the "better" of the two for the purposes of event history analysis.

2. Effects of Collection Method on Occupational Spells

The data employed in our analysis were extracted from the 1985 and 1986 SIPP Full Panel Longitudinal Research Files. The samples were limited to those individuals aged 15 years or older at the start of the panel who reported at

[1]This paper is based on research conducted under a Joint Statistical Agreement between the U.S. Bureau of the Census and the University of Toledo. The author would like to thank Dan Kasprzyk for suggesting this research topic. Rajendra Singh and Tom Scopp of the Census Bureau provided a number of helpful comments on earlier versions of this paper and Jim Lepkowski, Steve Pennel and David Miller of the University of Michigan assisted in obtaining the data. Any errors are the author's responsibility.

least two months employment in the January 1986 through April 1987 period. The data were "realigned" to calendar-month (as opposed to panel- month) observations on the basis of rotation-group membership prior to extracting the occupation, industry and wage measures.

Table 1 presents a variety of descriptive statistics for occupational spells by method of data collection for the 1985 and 1986 SIPP panels. The average number of occupational spells per individual observed in the 16 month observation period is some seventy percent higher for the independent data collection method (2.39) than for the dependent method (1.47). This is consistent with the higher amount of change noted in the introduction. Similarly, the amount of censoring is substantially higher with the dependent collection method than with the independent. More than two-thirds (66.8%) of the occupational spells observed in the 1986 panel (dependent collection) were left censored, as opposed to less than three-fifths (56%) in the 1985 panel (independent collection). Even more dramatic is the increase in right censoring and dual censoring (i.e. both left and right censoring) brought about by the dependent data collection method. Right-censored spells are nearly fifty-percent (65.1% vs. 43.9%) more prevalent with dependent data collection than with independent, while dual censored spells are more than twice as prevalent (43.8% vs. 21.3%). Thus, one effect of data collection method on event history analysis is to reduce the number of apparently informative (i.e. non-left censored) spells available for analysis--from 4867 for the 1985 Panel to 3578 for the 1986.

Table 1
Occupational Event History Statistics
by Mode of Collection

	Independent Mode (1985 Panel)	Dependent Mode (1986 Panel)
Sample Size (Weight-Sum in Thousands)	11,042 (107,766)	10,304 (96,115)
Average Number of Occupational Spells January 1986 - April 1987	2.39	1.47
Percent of Spells Left Censored	56.0	66.7
Percent of Spells Right Censored	43.9	65.1
Percent Left and Right Censored	21.3	43.8
Percent of Non-Left Censored Spells Beginning at Seam	82.0	52.5
Percent of Non-Right Censored Spells Ending at Seam	87.1	64.5
Percent of Completed Uncensored Spells Beginning and Ending at Seam	67.0	23.6

Whether or not the reduced number of non-left censored spells brought about by dependent data collection represents a loss of information is not clear, because we do not know what the true pattern of occupational change is. We do know, however, that in the absence of measurement error entrances and exits from occupations should be evenly spread over the months of the reference period. Furthermore, given the SIPP design we would expect approximately one-quarter of all occupational spells to begin (end) at a seam and only about a sixteenth to both begin and end at seam months. The extent of clustering of transitions at the seam months is, therefore, an indication of the extent of respondent errors in properly placing events in time. Table 1 indicates that, in this respect, the dependent data collection method provides cleaner data. Roughly eighty-five percent of all non-left censored spells observed with the independent collection mode ended at a seam month. This compares with less than sixty-four percent with the dependent collection mode. Even more revealing is the fact that more than two-thirds of the completed spells observed with the independent collection mode **both began and ended** at the seam. This compares with slightly less than a quarter of those observed with the dependent collection method.

Thus, while the number of seam-coincident occupation changes is higher than it should be for either mode of data collection, it is substantially closer to what it should be with the dependent mode.

3. Effects of Collection Method on Event History Model Estimates

While the above descriptive statistics suggest that the dependent data collection method results in less but cleaner data regarding the timing of entrances and exits from occupations, it does not necessarily follow that parameter estimates for event history models will be significantly affected by collection mode. To investigate this issue, we must first develop an explicit event history model and then compare the estimates obtained from the two collection methods. Because the appropriate treatment of left censored spells depends crucially on whether the hazard rate is a function of time in occupation or industry, we will include time explicitly in our formulation. Specifically, we will investigate models based on the following hazard function:

$$h(t_i) = \Lambda(\alpha + \beta' X_i + \gamma_1 t_i + \gamma_2 t_i^2 + \gamma_3 S(t_i))$$

where Λ is the logistic function ($\Lambda(z)=\exp(z)/(1+\exp(z))$), X_i is a vector of characteristics of individual i, β is a vector of effects of these characteristics on the hazard of exiting an occupation, t_i is the time the individual has been in the occupational spell, and $S(t_i)$ is a dummy variable equaling 1 if month t_i is a 'seam' month.

Allowing for right-censoring, the likelihood function for our model can be expressed as:

$$L = \prod_{t_i < t_{max}} f^i(t_i) \prod_{t_i \geq t_{max}} (1 - F^i(t_i))$$

where t_{max} is the right limit of the observation period, $f^i(t_i)$ is the probability density of individual i exiting at time t_i **given** that he/she has not exited prior to that time, and $F^i(t_i)$ is the corresponding cumulative density function. The first product in equation 2) represents the contribution to the likelihood function of the non-right censored spells, while the second portion represents that of the right censored spells. The cummulative density function is related to the hazard function by:

$$F(t_i) = \prod_{t=1}^{t_i - 1} (1 - h(t))$$

where h(t) is the hazard of exiting at time t. Similarly, the probability density function is related to the hazard function via:

$$f(t_i) = h(t_i) \prod_{t=1}^{t_i - 1} (1 - h(t))$$

Substituting equations 3) and 4) into equation 2) yields the following likelihood function for the discrete time model:

$$L = \prod_{t_i < t_{max}} h(t_i) \prod_{t=1}^{t_i - 1} (1 - h(t)) \prod_{t_i \geq t_{max}} \prod_{t=1}^{t_i - 1} (1 - h(t))$$

$$= \prod_{t_i < t_{max}} h(t_i) \prod_{i} \prod_{t=1}^{t_i^* - 1} (1 - h(t))$$

where $t^* = t_{max}$ for right censored cases.

When dealing with a Non-EPSEM sample such as the SIPP, one can apply sampling weights via the following weighted-likelihood function:

$$L_w = \prod_{t_i < t_{max}} h(t_i)^{w_i} \prod_{i=1}^{n} \prod_{t=1}^{t_{max}^* - 1} (1 - h(t))^{w_i}$$

where w_i is the individual's sampling weight scaled by the average weight of the sample.

Finally, equation 6) is made a function of α, β and γ by substituting equation 1) for $h(t_i)$, and consistent estimates of these parameters can be obtained by maximizing the natural logarithm of the result with respect to them.[1] While we shall maximize the logarithm of equation 6) directly using an algorithm written by the author, we should note that it is also possible to use packaged logit programs by creating t_i pseudo-observations for each of the i individuals in the sample (see e.g. Allison, 1984). Also, we should note that as the number of time periods in the observation period increases, the probability of an individual exiting in any one period decreases and that, in the limit, our model reduces to Cox's proportional hazards model.

4. Results

Table 2 presents the estimates obtained by maximizing equation 6) with respect to the parameters of equation 1) for the combined 1985-1986 SIPP panels as well as for each panel separately. The sample consists of the first non-left censored occupational spell observed for each individual.[2]

The question as to whether mode of data collection has a significant effect on event history analyses is a very straight forward one with a clear formal test procedure. Under the null hypothesis of no structural difference, the likelihood-ratio statistic $-2(\ln(L_c)-\ln(L_i) - \ln(L_d))$, where subscripts c, i and d represent combined, independent, and dependent data collection respectively, is distributed χ^2 with degree of freedom equal to the number of parameters in the model. In our case, this statistic is 1,178 with 12 degrees of freedom and the null hypothesis is clearly and soundly rejected. Thus, method of data collection makes a big and

very highly significant difference in the estimates obtained for event history analysis of occupational exits in the SIPP data. Which method yields the "better" estimates, however, is a question which is not so easily addressed. It requires examination of the individual estimated effects and some judgements regarding their "reasonableness".

Table 2
Discrete-Time Event History Analysis
Exit from First Non-Left Censored Occupation
(SRS t-ratios in parentheses)

	Combined Sample	Independent Mode (1985 Panel)	Dependent Mode (1986 Panel)
Constant	- 3.10** (-19.05)	- 3.38** (-15.84)	-1.37** (-4.76)
Time in Occupation	0.63** (20.55)	0.63** (14.42)	0.35** (6.93)
Time-Squared	- .08** (-26.12)	- .08** (-18.57)	- .06** (-10.39)
Whether Seam Month	3.00** (83.52)	3.58** (74.43)	1.94** (33.25)
Age (@start)	- .58** (-8.76)	- .52** (-6.12)	-1.04** (-8.45)
Age-Squared	.07** (8.16)	.06** (5.46)	.12** (7.35)
Wage	- .02** (-3.25)	- .03** (-3.07)	- .01** (-6.24)
Education	.04 (0.64)	.13 (1.55)	.01 (0.73)
Whether Black	.02 (0.44)	.05 (0.78)	- .01 (-0.11)
Whether Female	- .08* (-2.27)	- .11* (-2.50)	.02 (0.28)
Specific Vocational Prep.	- .05** (5.22)	- .03* (-2.34)	- .11** (-5.98)
Occ. Inconsistency	.05* (2.20)	- .01 (-0.39)	.14** (3.60)
Log Likelihood (base Log L)	-12,541 (-18,160)	-7,682 (-12,744)	-4,270 (-5,160)
Likelihood-Ratio Index (χ^2) df. = 11	30.88% (11,238**)	39.63% (10,123**)	17.04% (1,781**)
Number of Cases	10,372	6,798	3,574

**Significant at the .01 level.
*Significant at the .05 level.

The independent variables included in our model can be divided into three groups according to how firm our a priori knowledge is about their true effects on occupational exit hazards. The first group consists of whether the month in question is a seam month and the occupational coding inconsistency index.[3] **In the absence of measurement errors, these variables should have no effect on the exit hazards.** The second set of independent variables consist of time (and its square) in the occupation. While time in occupation should affect the exit hazard rate, our a priori's on the precise pattern of this effect are not very strong. The final set of independent variables are the substantive measures of characteristics of the individual (age, wage, education, gender) and the occupation (specific vocational preparation[4]). These variables should affect exit hazards, with higher hazards for younger low wage individuals in occupations with little specific human capital.

Perhaps the most important thing to note about the estimates of Table 2 is that overall goodness of fit of the model, as measured by the adjusted likelihood-ratio index, is substantially higher for the independent (39.6%) than for the dependent mode (17.0%). This is due almost entirely, however, to the gigantic effect of the seam with the independent mode data. The coefficient of 3.58, for whether the month in question is a seam month, implies that odds of exiting an occupation are some 35 times (=exp(3.58)) as high in seam months than in nonseam months. The corresponding effect for the dependent mode is just under 7 (exp(1.94)).

4.1 Seam and Inconsistency Index

The effect of this difference in the seam variable is that it dominates the model for the independent, but not for the dependent mode data. Table 3 presents the marginal adjusted likelihood-ratio indices[5] for the three sets of predictors by mode of collection. The explanatory power of the seam and inconsistency index with independent mode data collection (32.5%) is more than a full order of magnitude greater than the time variables (6.3%), and almost two orders of magnitude greater than the substantive variables (0.4%). While still the most important predictors with dependent data collection, the relative importance of the seam and inconsistency index is greatly reduced.

4.2 Time in Occupation

The positive and significant coefficient for time in occupation combined with the negative and significant coefficient for time-squared, for all three samples, indicates that the hazard of exiting an occupation increases at a decreasing rate with time for the first three or four months (i.e. the function $\gamma_1 t + \gamma_2 t^2$ attains a maximum at $t = \gamma_1/(2\gamma_2) = .63/(2*0.08) = 3.9$) and declines thereafter. The shape of the survival curve for the independent mode model is, by the way, virtually identical to that obtained by Hill and Hill, 1986, for unemployment exits using a Cox proportional hazards model on the 1984 SIPP panel.

4.3 Substantive Variables.

With respect to the substantive predictors, method of collection has strong impacts for estimated effects of some but not all measures. While the **direction** of the effects of age (at the beginning of the observation period), wage, and specific vocational preparation are the same for both data collection methods, their size and significance are much stronger for the dependent mode data. For both collection modes, occupational exit hazards decrease with age at a decreasing rate, until approximately age 45, at which point they begin rising again. This pattern is quite reasonable in that it reflects younger workers' higher risk of unemployment and older workers higher risk of leaving the labor force. The strength of these age effects for the dependent mode data is twice that of the independent. The estimated wage and specific vocational preparation effects are also consistent with our a priori knowledge and are much more highly significant, and in the case of the specific preparation, more powerful with the dependent mode data.

The only substantive variable which is more powerful for the independent data than for the dependent is gender. There is evidence that females have lower occupational exit hazards than men in the independent data. This result seems somewhat suspect. Historically, women workers had higher exit hazards than men--a reflection of more intermittent labor force participation. By 1986, work patterns had certainly changed, with women being more likely to work even when family demands peaked. However, whether these changes are enough to explain a reversal in exit hazards is doubtful.

Table 3
Explanatory Power by Variable Type and Mode
(Adj. Likelihood-ratio Index)

Variable Type	Independent Mode	Dependent Mode
Time	6.3%	4.1%
Seam	32.5%	11.4%
Substantive	0.4%	3.0%

Conclusions

In this paper, we have attempted to assess the relative quality of occupational data obtained from the independent and dependent modes by examining the association of measured occupational change with exogenous variables. This "empirical validity" was found to be significantly higher for the dependent (1986 SIPP Panel) mode data than for the independent (1985 SIPP Panel) mode data. Things which, in the absence of measurement error, should not effect occupational change (e.g., whether a seam month) had smaller effects with the dependent data, while things which should be associated with change had larger effects. These differences were very highly significant. Our analysis suggests that, at least within the context of event history models, the analytic potential of the SIPP occupation data was increased substantially by the move to dependent data collection methods.

References

Allison, P.D. (1984). Event History Analysis: Regression for Longitudinal Event Data. Newbury Park, Calif: Sage Publications.

Hill, D.H. (1992). Dependent and independent data collection in panel surveys: Analysis of 1985-1986 SIPP occupation and industry data. (memeo, Survey Research Institute).

Hill, D.H. & M.S. Hill (1986). Labor force transitions: Analysis of two panel surveys. Proceedings of the Section on Survey Research Methods. (pp. 220-225). Alexandria, VA. American Statistical Association.

Jabine, T.B.,& Tepping, B.J. (1973). Controlling the quality of occupation and industry data. The Bulletin of the International Statistical Institute, 45(3).

Miller, A.R., Treiman, D.J., Cain, P.S., & Roos, P.A. Ed. (1980). In National Research Council, Commitee on Occupational Classification and Analysis, Work, jobs and occupations: A critical review of the dictionary of occupational titles. Washington: National Academy Press.

---

1. The estimates are fully efficient only under the assumption of simple random sampling. The estimated sampling errors from maximizing equation 6) do not reflect the effects of departures from simple random sampling in the SIPP design and will tend to under state the true sample variability. Since the 1985 and 1986 SIPP designs are quite similar, however, comparisons of the estimated standard errors are still indicative of the relative precision of the estimates.

2. We discard left censored spells because we expect the hazard rate to be a function of time in occupation and these spells are uninformative. Non-left censored spells subsequent to the first are potentially informative but require one to assume independence between spells--a very strong assumption. If the independence assumption is violated, the use of multiple spells per individual will result in biased parameter estimates.

3. This index is taken from Jabine and Tepping's "Controlling the Quality of Occupation and Industry Data" and is the proportion of variation attributable to response error.

4. Specific vocational preparation is a measure of the amount of time required to become proficient in an occupation. It was merged via a 3-digit occupation code match with information published in **Work, Jobs and Occupations**, (Washington: National Academy of Sciences, 1980).

5. The marginal adjusted likelihood-ratio index (or marginal adjusted pseudo-R^2) is obtained via $\rho^2 = (L_u - L_r + k)/L_r$ where L_r is the log-likelihood value obtained when the 'k' coefficients relating to the variables under examination are restricted to 0 and L_u is the corresponding unrestricted log-likelihood value.