Industrial and Organizational Psychology (2019), **12**, 320–325
doi:10.1017/iop.2019.56

CAMBRIDGE UNIVERSITY PRESS

COMMENTARY

Turnover modeling and event history analysis

Rodney A. McCloy*, Justin D. Purl, and Erin S. Banjanovic

Human Resources Research Organization
*Corresponding author. Email: rmccloy@humrro.org

Speer, Dutta, Chen, and Trussell (2019) emphasized three words characteristic of the turnover field: "practical," "modeling," and "messier" (p. 277 [in the Abstract]), and advocated use of survival analysis (more generically, event history analysis, or EHA) as "a powerful tool for the purposes of turnover prediction [that] should be given consideration when planning turnover analyses" (p. 291). We agree and propose that this modeling approach best addresses the messy nature of practical turnover research.

Some researchers consider EHA impractical because the technique requires thinking about turnover in a more dynamic way than does the logistic "stay or go" approach. EHA, however, is more aligned with turnover theories (Hom, Mitchell, Lee, & Griffeth, 2012; Lee & Mitchell, 1994; Mobley, 1977), which imply a dynamic event, and it provides an analytic framework for properly modeling dynamic events such as turnover. Our commentary focuses on three primary topics: (a) some of the analytic challenges turnover data present, (b) a summary of EHA and how it allows researchers to meet those analytic challenges, and (c) a description of how to extend EHA to organizational decision making via transformation of EHA results to attrition probabilities for employees in your analysis sample and via decision support tools.

Analytic challenges of turnover data

Event data pose several analytic challenges (Allison, 1984; Singer & Willett, 1991, 2003). Consider a researcher interested in teacher turnover during the first 5 years of employment. Event data typically span a certain observation period (here, 5 years), during which time the units of analysis either do or do not experience the event in question (teachers either turnover or do not). During the observation period, it may be useful to measure objects in question on some of the predictor variables more than once. Hence, in addition to measures of the teacher's race, education, and age at hire, the researcher might wish to obtain biannual measures of each teacher's marital status during the 5-year observation period.

Three sticky analytic issues in this example are common to event data. First, there is the presence of time-varying independent variables (e.g., marital status). Conventional regression analyses do not handle time-varying measures well (Allison, 1984, pp. 10–11). Time-varying independent variables do not appear in every event study, but it is sensible to consider them, given that the observation of events naturally occurs over time, and a person's status on many potential predictors may be expected to change in a meaningful way during the elapsed time.

Second, some teachers will not leave the profession during the observation period. Yet other teachers might be lost to observation in the middle of the period (e.g., they switched to a school not included in the study), at which time they were still teaching. They did not turn over during the

The current affiliation for author Justin D. Purl is Google.

time they were observed, but they were not observed for the entire 5-year period. Observations that do not experience the target event are said to be "censored."

There are several types of censoring (cf. Lawless, 1982). The censoring in the example scenario is typically termed "right censoring." As time moves left to right on a timeline, a person's observation period runs along the timeline until it hits a border on the right side: the point where the observation period ends or where the person leaves the study. Any events or nonevents after (i.e., to the right of) this demarcation are unknown to the researcher (i.e., they are censored).

There are two types of right-censored observations in the data for the above scenario. First, those teachers who are teaching throughout the 5-year observation period are right censored at $t = 5$ years. Second, those teachers who exited the study for whatever reasons (e.g., they moved from the area and were lost to follow-up) are also right censored but at the point they exited the study.

The third analytic difficulty in the teacher turnover data regards the distribution of the events over time: They in no way approximate a normal distribution. This is the rule rather than the exception for event data. Hence, familiar analytic techniques based upon normal theory are inappropriate for event data.

A fourth analytic difficulty that is not present in the divorce example but is nearly ubiquitous in organizational research is *late entry* (also known as "left truncation"; Singer & Willett, 2003). Late entry occurs when the researcher cannot track every observation in a dataset from a common starting point (e.g., hiring date). Singer and Willett (2003) highlighted that this scenario occurs frequently with such *stock samples*: collections of people of different ages who already exist in the initial state of interest (here, employment) at the beginning of a research study (Lancaster, 1990).

Failure to account for late entry can lead to biased results. Consider a cohort of employees hired by Organization X in 2002. For a turnover study beginning in 2008, many in this cohort were already employed when the observation period began. They represent a restricted, unrepresentative sample of their employee cohort—a sample with a tendency to stay. Employees from this cohort who had a propensity toward turning over likely did so prior to 2008. We have little to no data on these individuals, and so they never enter our analysis. Failure to recognize we are dealing not with new employees from the 2002 cohort when our study begins in 2008 but rather with employees having 6 years of experience would lead us to underestimate turnover rates for the early years of employment. This example applies to all employees hired pre-2008.

Other factors requiring special attention also appear in event data, including the possibility of repeated events (e.g., absence-taking behavior, leaving then rejoining a company) or competing events (e.g., turning over for one reason rather than another). Fortunately, EHA allows researchers to adroitly handle all these difficulties.

An overview of EHA

Event history models involve functions, and the function values are evaluated across time. Time can be measured discretely (years, months) or continuously (days, hours, minutes). The two primary functions used in EHA are the survivor function and the hazard function.

The survivor function, $S(t)$, describes the probability an individual will survive at least until time t without experiencing the event in question. $S(t)$ is a monotonic, nonincreasing (typically decreasing) function. In this respect, it is essentially a reverse cumulative distribution function, cumulating across time the proportion of observations that have yet to experience the event.

Figure 1 presents sample survival curves for notional turnover behavior of men and women. The curves have been reflected to cumulate event occurrence over time (i.e., $1 - S(t)$). The curves show men and women take nearly the same amount of time to experience 25% turnover (4 years 8 months for men, 4 years 9 months for women), but 50% of men have separated nearly 2 years earlier (11 years 3 months) than 50% of women (13 years 2 months).

https://doi.org/10.1017/iop.2019.56 Published online by Cambridge University Press

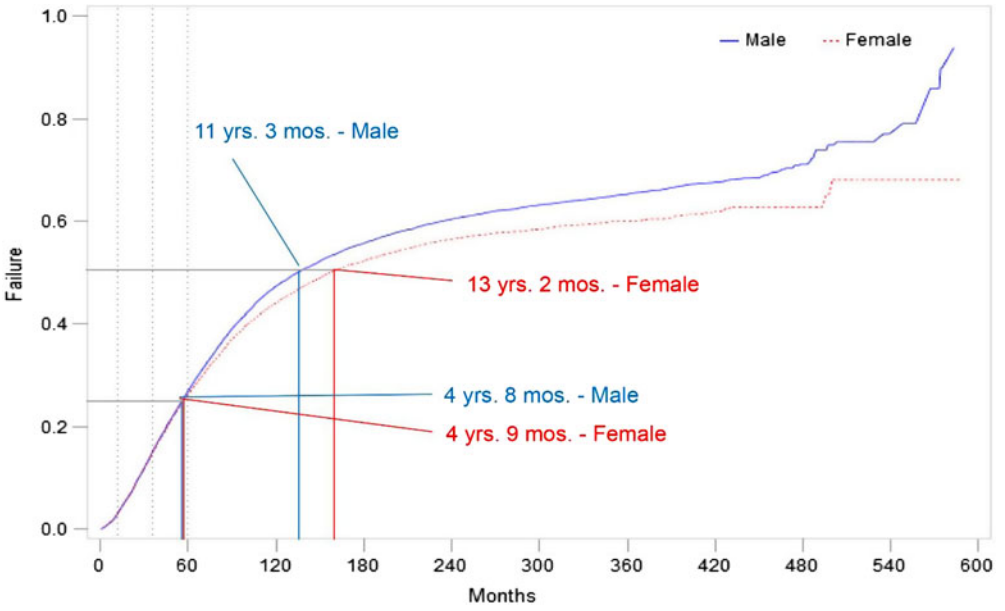

Figure 1. Notional inverse survival functions (1 – *S*(*t*)) highlighting time to 25% and 50% turnover for men and women.

The hazard function, $h(t)$, describes the distribution of event occurrence across time. The definition of $h(t)$ depends upon whether time is measured in discrete units (e.g., by month or year) or continuously.

Discrete time

For discrete time, $h(t)$ represents the probability an individual will experience an event during a particular time interval, given the individual is at risk for experiencing the event. Hence, the hazard is a conditional density function. Calculation of the discrete-time hazard depends on two quantities: (a) the number of individuals who experience the event during the interval, divided by (b) the number of individuals who are at risk for experiencing the event during the interval—what Allison (1984) labeled the *risk set*. For single events, the risk set steadily decreases as individuals either experience the event or are censored.

The risk set and its use in the calculation of the discrete time hazard demonstrate how EHA makes optimal use of data from censored observations. Referring again to the teacher turnover example, assume 500 teachers appear in the original sample at time $t = 0$ and observations of employment status are made biannually. At the first observation period of 6 months, say 15 of the 500 teachers will have stopped teaching. Hence, for the first 6-month period, the discrete time hazard is $15/500 = .03$. During the second time interval, 25 teachers leave the profession and 10 others exit the study while still teaching (i.e., they are censored observations). The risk set is now 485 rather than 500, because the 15 teachers who separated during the first 6 months are no longer part of the sample. Hence, $h(t)$ for the second time interval is $25/485 = .05$. Note that the censored observations contribute to the risk set for Interval 2 but are not considered events, because they did not stop teaching. For the third time interval, however, the risk set will be $(485 - 35) = 450$. Thus, the censored observations do not contribute to the risk set for the third interval. The data for censored observations are used correctly and optimally, contributing to the calculation of the hazard rate (via the risk set) for the amount of time the observations are in the study.

To control for late entry, we modify how we define the risk set for turnover in any given year (conditional partial likelihood; see Guo, 1993). For those hired in 2002 in our Organization X

https://doi.org/10.1017/iop.2019.56 Published online by Cambridge University Press

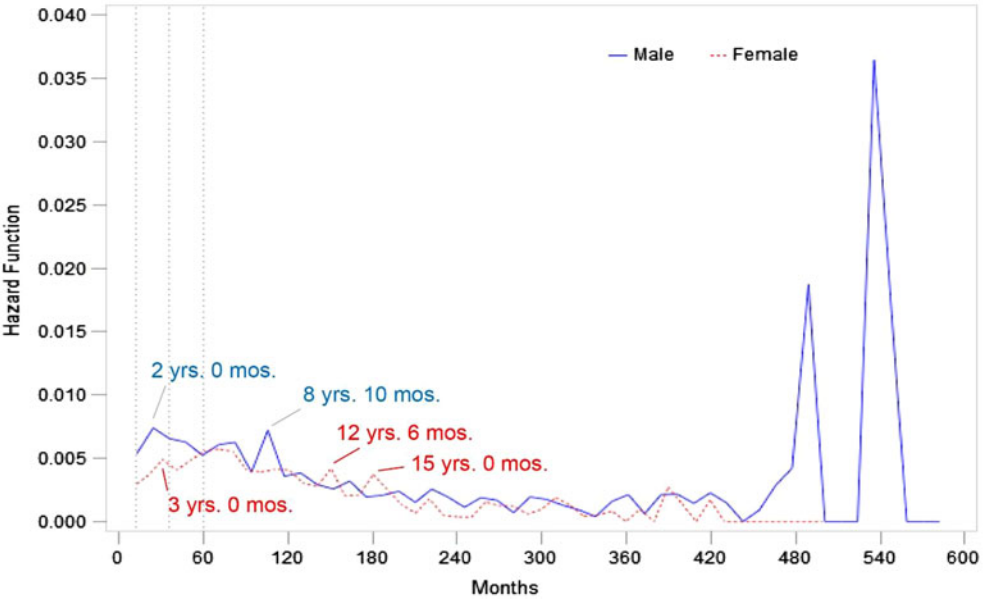

Figure 2. Notional hazard functions showing turnover rates for men and women.

example, their first opportunity to separate during our observation period occurs not in their first year of employment but rather in their sixth. Therefore, they are held out of the risk set in years 1 through 5, because they have already successfully navigated those years. They are, however, at risk for turnover from years 6 through 12 (2008–2014—the duration of our study) and thus appropriately enter the risk set at that time.

Continuous time

For continuous time, $h(t)$ is defined as a mathematical limit that describes the instantaneous rate at which events occur. Formally,

$$h(t) = \lim_{\Delta t \to 0} \frac{p(t \leq T < t + \Delta t | \, T \geq t)}{\Delta t}$$

where T is the time of the event (Kalbfleisch & Prentice, 1980, p. 6). The numerator is the discrete time hazard rate when $\Delta t = 1$. For the continuous time hazard rate, that probability is divided by Δt, the length of the interval; the limit is evaluated as this interval becomes smaller. The continuous time hazard is not a probability, because it can take on values greater than one (Allison, 1984).[1] The natural logarithm of the hazard rate typically serves as the dependent variable in event history models.

One way of conceptualizing the continuous-time hazard rate is to visualize a giant wheel attached to a large box full of people. As the wheel turns, people are ejected from the box. The people coming out of the box are those who experience the event in question. The faster the wheel turns, the faster people are thrown out of the box (i.e., the higher the hazard rate for the event).

Unlike the survivor function, $h(t)$ can take any shape. Figure 2 shows the hazard plots for the corresponding survivor functions in Figure 1. The turnover rate for men is highest at the 2-year

[1]The survival and hazard functions are related. Formally, the hazard rate is defined as $h(t) = \frac{f(t)}{S(t)}$ where $f(t)$ is the probability density function of the event times and $S(t)$ is the survivor function. Note that $f(t) = -\frac{dS(t)}{d(t)}$ because $S(t)$ is a reverse cumulative distribution function (Kalbfleisch & Prentice, 1980).

https://doi.org/10.1017/iop.2019.56 Published online by Cambridge University Press

point and then generally declines. The turnover rate for women increases after hire until its highest point around the 5-year mark.

There are many types of event history models, but a useful way to categorize them is by whether $h(t)$ defines events in discrete time or in continuous time. When time is modeled as discrete, $h(t)$ defines the conditional probability an individual will experience an event (turnover) during a particular time interval, given the individual is at risk for experiencing the event (a conditional density function). When time is modeled as continuous, $h(t)$ is the instantaneous rate at which events occur.

Discrete-time survival models can be estimated using conventional logistic regression packages run on a person-period dataset (i.e., a dataset where each row defines a unique person-time status, such as person's turnover status during month 1 in row 1, status during month 2 in row 2, etc.; see Singer & Willett, 1993, 2003). Continuous time survival models are most popularly represented by the proportional hazards model (Cox, 1972; Singer & Willett, 2003).

Extending EHA to organizational decision making

EHA makes the challenges of analyzing turnover data explicit and thus seems more complicated at first glance. However, by focusing on the survival function, EHA results are no more complicated to interpret than logistic regression results. Specifically, weights for each predictor reflect a relative strength of the driver of turnover, and combinations of predictor standings can be transformed into intuitive probabilities.

To begin, the survival function for individual i at time j is

$$S(t_{ij}) = S_0(t_j)^{e^{(\beta_1 X_{1i} + \beta_2 X_{2i} + \ldots \beta_P X_{Pi})}}$$

(Singer & Willett, 2003, pp. 532, 540). This function has two components: (a) $S_0(t_j)$ the baseline survival function, where all predictors (X_{1i}, X_{2i}, $\ldots$, X_{Pi}, where i = individuals 1 to N) are set to zero; and (b) the risk score, $e^{(\beta_1 X_{1i} + \beta_2 X_{2i} + \ldots \beta_P X_{Pi})}$, a transformation of the linear combination of the predictors.

The survival function at a particular time point is conceptually similar to a predicted probability in the logistic regression framework. At a particular time (t_j), the baseline survival function becomes a constant, greatly simplifying the above function. Note that the outcome of this model (survival or event non-occurrence) is the inverse of a typical logistic model (event occurrence). To get the latter, we compute the inverse of survival, 1 – $S(t)$ (as in Figure 1), which may be termed failure (F). Our final model to capture the logistic equivalent of the survival analysis at a particular time point then becomes

$$S_i = S_0^{e^{(\beta_1 X_{1i} + \beta_2 X_{2i} + \ldots \beta_P X_{Pi})}}$$

$$F_i = 1 - S_i$$

If we select a time that corresponds to the timeframe of a logistic regression analysis, the two probabilities reflect similar ideas: the probability an individual will leave before the chosen timepoint. Whereas this could be an oversimplification of the results for those intimately familiar with EHA, the logistic framing of survival results could be a beneficial way of explaining the analysis to a broader audience.

A particularly effective way to present results using predictive model weights is through decision support tools (Diaz et al., 2014), which allow leader input to be translated through analysis results (via simulation) into numerical or graphical presentations of hypothetical scenarios. The "what-if" process clarifies the population level results for choices such as investing in increasing a particular organizational unit's standing on a key driver of turnover. A decision support tool can estimate the reduction in number of employees leaving the organization from that unit over a given period assuming satisfaction is increased by 0.5, 1.0, or 1.5 units. Knowing the cost of

https://doi.org/10.1017/iop.2019.56 Published online by Cambridge University Press

turnover in the given organizational unit (say $60,000/employee), return on investment (ROI) can be calculated for an increase of each hypothetical level. If the investment will not pay off even with a large increase in satisfaction, the organization can apply those resources elsewhere.

Similarly, simulations can serve as a workforce planning tool. Instead of hypothetical scenarios, new employees can be added to the simulation as they arrive, and failure for all employees can be calculated at several timepoints in the future. The support tool provides an expected drop out at each quarter, allowing the organization to plan compensatory recruitment activities. By adding predictors to the analysis behind the support tool that reflect market-level trends, the workforce planning approach can be tied to changes expected in the economy or industry for the next quarter.

Workforce planning and ROI analysis can also be powered by linear, logistic, or multilevel regression. Decisions based on analytic tools, however, are subject to the error in the underlying models. Using inappropriate models will introduce error in parameter estimates that is unknowable to the researcher. As stewards of the analyses underlying organizational decision making, we are obligated to use the best techniques available to support decision makers. In the case of turnover, we believe EHA is that technique and that the additional complications of modeling with EHA are surmountable.

References

Allison, P. D. (1984). *Event history analysis: Regression for longitudinal event data* (Sage University paper series on quantitative applications in the social sciences, Number 07-046). Beverly Hills, CA: Sage.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society*, **34**, 187–202.

Diaz, T. E., Sticha, P. J., Mackin, P., Hogan, P., Rinde, S., & Jose, I. (2014). *Decision support tool prototype for the Enlistment Incentive Review Board: Phase 2 (interim report)*. Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Guo, G. (1993). Event-history analysis for left-truncated data. *Sociological Methodology*, **23**, 217–243.

Hom, P. W., Mitchell, T. R., Lee, T. W., & Griffeth, R. W. (2012). Reviewing employee turnover: Focusing on proximal withdrawal states and an expanded criterion. *Psychological Bulletin*, **138**(5), 831–858.

Kalbfleisch, J. D., & Prentice, R. L. (1980). *The statistical analysis of failure time data*. New York, NY: John Wiley and Sons.

Lancaster, T. (1990). *The econometric analysis of transition data*. New York, NY: Cambridge University Press.

Lawless, J. F. (1982). *Statistical models and methods for lifetime data*. New York, NY: John Wiley & Sons.

Lee, T. W., & Mitchell, T. R. (1994). An alternative approach: The unfolding model of voluntary employee turnover. *Academy of Management Review*, **19**, 51–89.

Mobley, W. H. (1977). Intermediate linkages in the relationship between job satisfaction and employee turnover. *Journal of Applied Psychology*, **62**(2), 237–240.

Singer, J. D., & Willett, J. B. (1991). Modeling the days of our lives: Using survival analysis when designing and analyzing longitudinal studies of duration and the timing of events. *Psychological Bulletin*, **110**(2), 268–290.

Singer, J. D., & Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics*, **18**(2), 155–195. doi:10.3102/10769986018002155

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis*. New York, NY: Oxford University Press.

Speer, A. B., Dutta, S., Chen, M., & Trussell, G. (2019). Here to stay or go? Connecting turnover research to applied attrition modeling. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, **12**(3), 277–301.

Cite this article: McCloy, R.A., Purl, J.D., and Banjanovic, E.S. (2019). Turnover modeling and event history analysis. *Industrial and Organizational Psychology* **12**, 320–325. https://doi.org/10.1017/iop.2019.56