Paper 070-31

An application of Survival Analysis to Population Dynamics in Human Capital Management

Martin Jetton/Dr Robert Yerex, Unicru, Inc, Beaverton, OR

ABSTRACT

Using Base SAS®, SAS/STAT® and SAS/GRAPH® we have built a decision support infrastructure to monitor and forecast workforce retention dynamics relative to the economic value of Unicru Personality Assessments. Using PROC LIFEREG® we develop best fit survival models for employee lengths of stay and with these best fit model's hazard rates and apply a semi-Markov population dynamic model (described in this paper) to forecast the rate of replacement of employees with new hires with targeted personality traits. SAS/GRAPH® is utilized to present historical data, fitted survival models and forecasted replacement levels.

INTRODUCTION

Unicru Inc. provides many retail customers with personality assessments to target specific traits in the hiring process in an hourly workforce. These personality assessments address specific strategic issues such as improved sales, customer service, time and attendance behavior and productivity. We help our customers understand the economic impact and project future benefits of Unicru's personality assessments by applying a population dynamics model based upon employment spell hazard rates.

Many measurable factors within the work environment have an impact on, or can be related to, employee asset value. Unicru's Incremental Employee Contribution or IEC model has been developed to help organizations address key questions that relate to the value of their employees, and to examine alternative interventions aimed at optimizing the asset value of their employee pool. To date, the IEC model has been used successfully to examine relationships between length of service, time to competency, productivity, the value of retention, employee asset value, and ultimately employee profitability.

In the examination of alternative interventions for optimizing employee asset value, a logical intervention is to select for effecting behaviors at the time of the hiring decision. Unicru's personality assessments provide an effective intervention mechanism for organizations to target effecting behaviors in the hiring process. With each new hire selected using Unicru's personality assessments, the number of employees predicted to display the effecting behavior increases. The rate at which this selection pressure impacts the employee pool can be forecasted using two aspects of the hiring process; the rate at which employees are retained and compliance in hiring employees with the targeted behavioral trait. The portion of the employee pool (or headcount) resulting from applying selection pressure using the Unicru System is referred to as the Unicru Replacement Level. For the IEC model we are interested in calculating the Unicru Replacement Level at a point in time from Unicru implementation given different retention and compliance rates.

Replacement rate estimation is, in practice, much more complex than it may initially appear to be. The stochastic and semi-deterministic factors lead to difficulty in creating useable closed form solution estimators. Simulation techniques are often used to control for these factors. In practice such simulation techniques can lead to reasonably accurate predictions, but they are often time consuming and cumbersome to carry out. The closed form estimator of Unicru Replacement level developed in this paper is practical and reasonably accurate when compared to actual and simulation results.

For further discussion of the IEC see the white paper "Human Capital Management – Establishing a Meaningful Metric Framework", by Dr Robert Yerex. This white paper is available from Unicru, Inc at www.unicru.com.

VALUE OF UNICRU REPLACEMENT LEVEL

The success in applying Unicru assessment strategies is dependant upon the rate at which employees with the effecting behavior(s) replace headcount in the employee pool. The following example illustrates the rate of replacement and the time it takes to build value associated with the Unicru Sales Assessment.

Example: A customer signs up to use the Unicru Sales Assessment. Their goal is to achieve an improvement in sales productivity for commissioned sales personnel. In comparing this customer to similar Unicru customers it is expected that the Unicru Sales Assessment selects hires providing 2% more sales per hour than non-Unicru hires.

Assuming that both Unicru and non-Unicru exhibit 100% turnover rates and a 100% selection pressure compliance, we would expect the sales per hour to reflect a 1% increase at 253 days (½ of the 2%). We would expect an aggregate increase in sales to reach 2% some 4 years after adoption of the Unicru process. If the customer experiences an 80% compliance rate, the customer would not realize the 1% sales per hour impact until one full year after introducing Unicru. And with the 80% compliance rate, the customer's maximum benefit would be 80% of 2% or 1.6% sales per hour. As you can see it takes time for this customer to realize the potential of Unicru Sales Assessment selection pressure impact on sales per hour. The example also illustrates how reduced compliance reduces the selection pressure and the long run impact the desired goal of an increase in sales per hour. A closed form solution to the Unicru Replacement Level will help customers understand that the impact of selection pressure takes time.

UNICRU REPLACEMENT LEVEL

Underlying Unicru Replacement Level are two fundamental drivers, compliance in hiring individuals with the affecting personality trait and the rate at which employees are retained. The first driver, selection pressure compliance, is easy to understand and easy to measure. Compliance level is the rate at which new hires are indicated as "Green" on Unicru personality assessments for specific effecting behaviors of interest. Selection pressure compliance is measured using payroll hires compared to applicants scoring green on assessments of interest.

Retention, the second driver of Unicru Replacement Level, relates to how long people stay. Often confused with turnover rates, retention is the measure of the length of time people are employed. Turnover measures the number of separations. Turnover does not measure the length of time that people stay with an employer. Retention analysis provides for the application of far richer analytical techniques by focusing on the length of stay. Underlying these analytical techniques is the concept of survival analysis and hazard rates, or the probability of exit of an employee given their length of stay.

RETENTION MODELING

Modeling lengths of stay or retention is critical in the development of a closed form estimator of Unicru Replacement Level. We define retention as the length of continuous employment from hire date to separation date. We'll think of this length of stay as a continuous random variable, **S**, and consider a large population of people who hired at a time **S**=0. S does not refer to calendar time but rather it measures a time on a person-specific clock that are each set to zero at the moment a person is hired. **S** is the duration of stay as a hire. The population is assumed to be homogenous with respect to the systematic factors, regressor variables that affect the distribution of **S**. This means that everyone's duration of stay will be a realization of a random variable from the same probability distribution. This length of time **S**, or duration, can be understood with the following three functions of time:

1)	Distribution Function	$F(s) = Prob(S < s)$
2)	Survival Function	$S(s) = Prob(S > s) = 1 – F(s)$
3)	Density Function	$f(s) = dF(s)/ds = -dS(s)/ds$

The distribution function, F(s), is defined as the probability that a hire will stay at least s days. The survival function, S(s), is defined as the probability that a hire will stay more than s days. The density function, f(s), is the probability that a hire will stay exactly s days.

For the modeling of Replacement we need to know, for a given length of stay s, what is the conditional probability of exit during the next increment of time (or delta):

$$h(s, \Delta) = Prob\ (s < S < s+\Delta \mid S>=s)$$

The rate of exit or hazard rate is defined as:

$$h(s) = \lim_{\Delta \to 0} h(s,\Delta)/\Delta = f(s) / S(s) \qquad (1)$$

The hazard rate represents the instantaneous probability that the employee separates at time s, conditional on the fact that they have lasted up to time s. The hazard rate is equal to the probability a hire will stay exactly s days divided by the probability a hire will stay more than s days.

In practice, to measure hazard rates we review historical lengths of stay equal to separation date minus hire date. This length of stay is used to develop a density function of durations of stay, s. The density function, f(s), is the

probability that a hire will stay exactly s days. From the density function, the distribution function and survival functions can easily be calculated. The hazard rate is the density function divided by the survival function.

With system compliance and hazard rates, or the probability of exit of an employee given their length of stay, we develop the closed form estimator of Unicru Replacement Level.

UNICRU REPLACEMENT LEVEL

The Unicru Replacement Level at time s from adoption of Unicru is given by the formula:

$$UR(s) = \frac{P_{N->U}}{P_X} - \frac{P_{N->U}}{P_X} e^{-P_X}$$

e^x is the natural log.

The closed form estimator of Unicru Replacement Level is composed of two parts. First $\frac{P_{N->U}}{P_X}$, is the rate at which 'non-compliant' hires become Unicru compliant as percent of total hires. The second, $\frac{P_{N->U}}{P_X} e^{-P_X}$, provides that the rate us based upon length of time from Unicru adoption. The function, e^{-P_X}, goes to 0 (zero) as time, s, increase and thus UR(s) approaches $\frac{P_{N->U}}{P_X}$.

The terms $P_{N->U}$ and PX are developed from the cumulative hazard rate, compliance rate and time from Unicru adoption. Using the cumulative hazard multiplied times the compliance rate, we calculate the probability of a position alternating from a non-Unicru hire to Unicru hire:

$$P_{N->U} = \text{Compliance} * CH(s)_N$$

where $CH(s)_N$ is the Cumulative hazard for non-Unicru Hires. The probability of exit of a non-Unicru hire over a period of time.

The probability of a non-compliant separation is defined as the probability of alternating between non-Unicru and Unicru and the probability of alternating between Unicru and non-Unicru.

$$P_X = P_{N->U} + P_{U->N}$$

where $P_{U->N}$ = (1-Compliance) * $CH(s)_U$. The probability of a position alternating from a Unicru hire to a non-Unicru hire up to time s. $CH(s)_U$ is the Cumulative Hazard for Unicru Hires. The probability of exit of a Unicru hire up to time s. The closed form estimator of Unicru Replacement Level represents the rate of change of a position from a non-Unicru hire to a Unicru hire, $\frac{P_{N->U}}{P_X}$, adjusted over time for non-compliance, $\frac{P_{N->U}}{P_X} e^{-P_X}$. See appendix A for the development of UR(s).

SAS IMPLEMENTATION

To implement the tracking of and forecasting for Unicru's Replacement Level, we need to identify the hazard rate functions for a new or prospective customer. To do this we use survival analysis and the procedure available SAS/STAT®, PROC LIFEREG®. At Unicru, as an ASP (application software provider), we collect total hires and separations at our customers for reporting back to them retention data. We (Unicru) and our customers use this data to evaluate the impact of personality assessments on retention. Given that we have ongoing interactions with our current customers payroll data (where the hires and terminations are captured) we find it easy to work with prospects to evaluate their historical retention from payroll data. Using our regular analysis of current customers, we have found the best distributions for employment spell data are the exponential, Weibull, log-logistic and log-normal. Consistently these four distributions return the best fit of the available distributions in PROC LIFEREG®. For purposes of our estimation of new customer retention behavior we focus our efforts on these four parametric functions.

Extensive discussion of survival analysis and SAS can be found in "Survival Analysis using SAS, A Practical Guide" by Paul Allison. Paul Allison outlines survival analysis and the fitted hazard functions we needed for the Unicru Replacement Level described above.

For retention analysis we use the difference between hire date and termination date as the duration and consider censored observations as those still employed at the time of the analysis. One of the advantages that PROC

LIFEREG provides is the handling of the random censoring nature of employment spell data. Employees are starting and ending at random points over time.

IDENTIFYING THE BEST FIT SURVIVAL MODEL

The first step is isolating the data for the time frame of interest. We use two macro variables 'max_obs_date' and 'min_obs_date' to select the date range of interest. In the first data step we calculate the duration and whether the observation was censored. Individuals who survived past the max date are considered censored.

Source code:

```
%let Mdlofint=;
%let Unicru_impact = 1.0;
%let Unicru_compliance = 1.0;
%let max_obs_date = %sysfunc(mdy(6,30,2005));
%let min_obs_date = %sysfunc(mdy(7,1,2003));

DATA work.SURVIVAL_DATA; set mywork.infoallemployees_spec;
*if uhirejobcodedescription in ("Clerk","Cashier");
if termdate_sas <= &max_obs_date then end_date = termdate_sas;    else
end_date = &max_obs_date;
if hiredate_sas <= &min_obs_date then delete;

DUR = end_date - hiredate_sas;   if DUR>0;
if termdate_missing = 1 or hiredate_missing = 1  then STATUS=0; else STATUS=1;
if termdate_sas >= &max_obs_date then status=0;

KEEP DUR STATUS HIREMTHYR hireyear hiremonth hiredate_SAS exposure tenure
employee;
RUN;
```

To capture the parameters provided in the modeling output we utilize SAS/ODS®. At the same time we are capturing the output from the modeling process for review.

```
ODS listing close; ODS pdf file='pdflifereg.pdf';
ODS output "Analysis of Parameter Estimates" (MATCH_ALL=parmsall
PERSIST=PROC)=ParmEstimates
           "Model Information" (MATCH_ALL=modelsall PERSIST=PROC)=ModelInfo;

PROC LIFEREG data=SURVIVAL_DATA ; title 'Log Normal';
      model dur*status(0) = &Mdlofint / DIST=Lnormal; RUN;
PROC LIFEREG data=SURVIVAL_DATA ; title 'Log Logistic';
      model dur*status(0) = &Mdlofint / DIST=llogistic; RUN;
PROC LIFEREG data=SURVIVAL_DATA ; title 'Exponential';
      model dur*status(0) = &Mdlofint / DIST=exponential; RUN;
PROC LIFEREG data=SURVIVAL_DATA ; title 'Weibull';
      model dur*status(0) = &Mdlofint / DIST=weibull ; RUN;
ODS output close;
```

After closing the ODS OUTPUT we have captured the model information to evaluate the best fit model. We then print the sorted models by the highest log likelihood.

```
DATA allmodels (keep=depndvar distrib loglikeli) ;
set &modelsall;
length depndvar $15.;
length distrib $15.;
retain depndvar distrib loglikeli;
if Label1 eq "Dependent Variable" then depndvar = cValue1;
if Label1 eq "Name of Distribution" then distrib = cValue1;
```

```
if Label1 eq "Log Likelihood" then do;    loglikeli = nValue1;
      output allmodels; end;
      else delete;
RUN;

PROC SORT data=allmodels; by descending loglikeli; run;
PROC PRINT data=allmodels;
  title 'Best Models Sorted by Log Likelihood'; run;
ods pdf close; ods listing;
```

IDENTIFYING THE BEST FIT SURVIVAL MODEL PARAMETERS

The ODS OUTPUT statement is used to capture the parameter estimates in 'parmestimates' data sets. Here we recombine them and re-label to select the best fit parameters to be used in estimating the survival functions. The best fit model parameters are selected into macro variables for use in forecasting.

Source code:

```
PROC TRANSPOSE data=parmestimates out=parms; var estimate; run;
PROC TRANSPOSE data=parmestimates1 out=parms1; var estimate; run;
PROC TRANSPOSE data=parmestimates2 out=parms2; var estimate; run;
PROC TRANSPOSE data=parmestimates3 out=parms3; var estimate; run;

DATA testparms;  set parms(in=inlognorm)
        parms1(in=inloglog)
        parms2(in=inexpon)
        parms3(in=inweibull);
length distrib $15;
if inlognorm then distrib ="Lognormal";
if inloglog then distrib ="LLogistic";
if inexpon then distrib ="Exponential";
if inweibull then distrib ="Weibull";
rename col1=Intercept;
rename col2=Scale;
rename col3=WiebullScale;
rename col4=WiebullShape;
   RUN;

PROC SQL noprint;  select distrib into :keydistrib
 from allmodels having loglikeli = (select max(loglikeli) from allmodels);
 select intercept,scale, WiebullScale, WiebullShape
      into :intercept,:scale,:WiebullScale,:WiebullShape
 from testparms where distrib = left("&keydistrib");
 QUIT;
```

ESTIMATING SURVIVAL FUNCTIONS AND GRAPHING

Now that we have captured the parameters and the best fit model we can use the parametric definition of the models to create a forecast of replacement levels based upon survival functions. The first DATA step creates a replacement curve for each 7 days (a week) increment out to 450 days after the implementation of Unicru. The replacement levels are set based upon the identified best fit survival functions. The remainder of this code compares the observed density distribution and hazard function to the fitted functions used PROC SQL® to summarize the data and SAS/GRAPH's ® PROC GPLOT® to plot the functions for evaluation.

Source code:

```
%let max_survival_days=450;
%let survival_incr=7;

DATA survival_theo (drop=alpha gamma); survival_days=1;
```

```
currdistrib = "&keydistrib";  cummhazrd_unicru  = 0;
 cummhazrd_nonunicru  = 0;
do while (survival_days <= &max_survival_days ) ;
 if currdistrib eq 'Lognormal' then do;
      alpha = 1/(sqrt(2*constant('pi'))*&scale*survival_days);
      gamma = exp(-.5*((log(survival_days)-&intercept)/&scale)**2);     surv =
1-probnorm((log(survival_days)-&intercept)/&scale);
      func = &survival_incr * alpha * gamma ;
      hazrd = func / surv;          end;
 if currdistrib eq 'LLogistic' then do;
            alpha = exp(-&intercept/&scale);
            gamma = 1/&scale;
            surv = 1/(1+alpha*(survival_days**gamma)) ;
            func = &survival_incr * alpha*gamma*(survival_days**(gamma-1))
/((1+alpha*(survival_days**gamma))**2) ;
            hazrd = func / surv;                end;
 if currdistrib eq 'Weibull' then do;
            alpha = exp(-&intercept/&scale);
            gamma = 1/&scale;
            surv =exp(-alpha*(survival_days**gamma));
            func =  &survival_incr *
      gamma*alpha*(survival_days**(gamma-1)) * exp(-
      alpha*(survival_days**gamma));
            hazrd =   func / surv;              end;
 if currdistrib eq 'Exponential' then do;
         alpha = exp(-&intercept);
            surv = exp(-alpha*survival_days) ;
            func = &survival_incr*alpha* exp(-alpha*survival_days)  ;
            hazrd =   func / surv; end;
 hazrd_nonunicru = hazrd;
 hazrd_unicru =  hazrd_nonunicru * &Unicru_impact;
 cummhazrd_unicru = cummhazrd_unicru + (1 - &Unicru_compliance) *hazrd_unicru
;
 cummhazrd_nonunicru = cummhazrd_nonunicru + &Unicru_compliance *
hazrd_nonunicru;
 cummhazrd = cummhazrd_unicru +cummhazrd_nonunicru;
 Uni_replace = (cummhazrd_nonunicru / cummhazrd ) - ( cummhazrd_nonunicru /
cummhazrd ) * exp(-cummhazrd);
 min_exposure = survival_days;
 max_exposure = survival_days + &survival_incr;
 output;
 survival_days = survival_days + &survival_incr;
 end;
RUN;

PROC GPLOT data=survival_theo;
title ' Unicru Replacement vs Theoretical Survival';
plot uni_replace*survival_days   surv*survival_days / overlay legend;
RUN; QUIT;

PROC SQL; create table sumexposure  as
   select s.survival_days, sum(employee) as emps_exposed
   from survival_theo s, survival_data a
   where a.exposure >= s.min_exposure and hiredate_SAS >= &min_obs_date
      and hiredate_SAS <= &max_obs_date
   group by s.survival_days; QUIT;

PROC SQL; create table sumtenure  as
   select s.survival_days, sum(employee) as emps_severed
```

```
   from survival_theo s, survival_data a
   where a.tenure >= s.min_exposure and a.tenure < s.max_exposure
   and hiredate_SAS > &min_obs_date  and  hiredate_SAS <= &max_obs_date
   group by s.survival_days; QUIT;

PROC SQL data=sumexposure; by survival_days ; RUN;
PROC SORT data=sumtenure; by survival_days ; RUN;

DATA density_dist;
merge sumexposure sumtenure;
 by survival_days ;
 if emps_exposed<0 then  emps_exposed=0;
 if emps_severed<0 then  emps_severed=0;
 if emps_exposed>0 then density_dist= emps_severed / emps_exposed;      else
density_dist= 0;
retain cumm_density_dist 0;
if _n_=1 then cumm_density_dist=0;  else
cumm_density_dist=cumm_density_dist+density_dist;
      surv_density_dist = 1 - cumm_density_dist;
      hazard_dist = density_dist / surv_density_dist;
RUN;

DATA survival_actual;merge density_dist survival_theo;by survival_days; RUN;

PROC GPLOT data=survival_actual;
title ' Survival Function (Actual vs Theo.)';
plot surv_density_dist * survival_days
     surv * survival_days /overlay legend;
       RUN;QUIT;

PROC GPLOT data=survival_actual;
title ' Hazard Function';
plot hazard_dist * survival_days hazrd * survival_days /overlay legend;
       RUN; QUIT;

PROC GPLOT data=survival_actual;
title ' Density Function';
plot density_dist * survival_days func * survival_days /overlay legend;
       RUN;  QUIT;
```

ACTUAL REPLACEMENT COMPARISON

Once a customer has implemented Unicru we can track the estimate to actual replacement levels. The following code selects observations relevant to post-Unicru implementation and compares to the fitted replacement levels out to 450 days.

Source code:

```
DATA replace_DATA;set mywork.infoallemployees_spec;
keep employee unicruapp hiredate_sas termdate_sas;
RUN;

DATA replace_days; survival_days=1;
 do while (survival_days < &max_survival_days);
   min_date = survival_days + &max_obs_date;
   max_date = min_date + &survival_incr;
   output;
   survival_days = survival_days + &survival_incr;
```

```
  end;

PROC SQL; create table sum_unicru  as
   select s.survival_days, sum(employee) as unicru_hires
   from replace_days s, replace_data a
   where hiredate_SAS <= max_date  and  termdate_SAS > min_date and
    unicruapp="Y" and hiredate_SAS <=&max_obs_date+&max_survival_days
   group by s.survival_days;RUN;

PROC SQL; create table sum_nonunicru  as
   select s.survival_days, sum(employee) as nonunicru_hires
   from replace_days s, replace_data a
   where hiredate_SAS <= max_date  and  termdate_SAS > min_date
      and unicruapp="N" and hiredate_SAS<=
            &max_obs_date+&max_survival_days
   group by s.survival_days;QUIT;

DATA compare_replace;
 merge survival_theo (keep=survival_days Uni_replace)
         sum_unicru
         sum_nonunicru;
 by survival_days;
 if nonunicru_hires < 0 then nonunicru_hires = 0 ;
 if unicru_hires < 0 then unicru_hires = 0 ;
 ttl_headcount = nonunicru_hires + unicru_hires;
 Act_Replace =  unicru_hires / ttl_headcount;
      RUN;

PROC GPLOT data=compare_replace;
 title 'Actual vs Predicted';
  plot Act_Replace * survival_days
         Uni_Replace * survival_days / overlay legend;
 RUN;
 QUIT;
```

CONCLUSION

We run into issues that dramatically impact our ability to forecast Replacement rates and levels. Primary among these problems are:

1) Rollout at customers takes time. With customers having many sites to implement the Unicru system for hiring, the lag effect of implementation can take several months. Also, when customers run tests for significant periods of time, these stores/sites need to be removed from the analysis.
2) Seasonality. Customers with significant seasonality will impact the prediction of retention. We attempt to identify previous hires who may have been seasonal hires through the data provided by customers. We exclude these hires from the analysis of history.
3) Diversity of positions. We know from experience that the attributes of positions such as full time/part time, regular/temporary (seasonal) and job categories such as manager, clerks or cashiers, will impact retention. We will split these into different retention analysis as needed.

REFERENCES

Allison, Paul, 1995, *Survival Analysis using SAS, A Practical Guide,* Cary, NC: SAS Institute Inc.

RECOMMENDED READING

Lancaster, Tony, *The Econometric Analysis of Transition Data*, Cambridge University Press, 1990 (1992 paperback printing), New York, NY, USA.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Martin Jetton, Senior Analytic Consultant,
Unicru, Inc
955 SW Gemini Dr
Beaverton, OR 97008
Work Phone: 503-596-3181
E-mail: mjetton@unicru.com
Web: www.unicru.com

Dr Robert Yerex, Director, Analytic Research Group
Unicru, Inc
955 SW Gemini Dr
Beaverton, OR 97008
Work Phone: 503-596-3181
E-mail: mjetton@unicru.com
Web: www.unicru.com