

SURVIVAL ANALYSIS WITH SAS LIFETEST AND PHREG PROCEDURES*

Zhongmin Li

for Poster Session of Third Annual South-Central Regional
SAS Users Group Conference

Introduction

Researchers conducting a survival analysis face the challenge of choosing a competent software program to handle the censored data and produce meaningful statistics. Two programs, SAS LIFETEST and PHREG, are notable for their ease of use and flexibility.

This paper illustrates the application of these two SAS procedures and provides opportunity for discussion and exchange of knowledge in survival analysis techniques and programming.

The sample application compares the outcomes under two approaches, treatments A and B. The study is designed to answer three related but distinct questions:

1. Has the outcome changed from treatment A to B?
2. To what extent have social-demographic factors affected the outcome?
3. Once the significant factors are found, what would happen to the outcome if these factors were changed under treatment B?

Each question requires a specific analytical technique. For the first question, a stratified survival analysis is used to produce descriptive statistics on the outcomes under treatments A and B. For the second question, a multiple regression is indicated. This study uses the Cox proportional hazard model. The third question is addressed using the results from the regression analysis of question two in cooperation with a hypothetical baseline data to predict the outcome for each given scenario under treatment B.

Data Preparation

Because data under treatments A and B is maintained in two separate data bases, the first step in data preparation is to create a combined sample data. A scientific sampling technique is used to randomly select 2,000 cases under treatment A and 2,500 cases under treatment B. The selected cases create a combined sample data set¹.

Figure 1 presents the coverage of this combined sample data, which can be divided into two types of cases: censored and noncensored. Due to the dynamic nature of case

* Author's note: I thank Vicky Hill for her valuable editing of an earlier draft.

¹ For random sampling with SAS-AF, see Zhongmin Li, "SAS-AF Application in Determining Sample Size," *Proceedings of the South-Central Regional SAS Conference '92*.

processing, it sometimes takes a number of years to obtain an outcome. To ensure that all cases in the study will be weighted comparably, an identical time span is assigned to both treatment A and treatment B.

If closure has not occurred by the cutoff date, the observation will be identified as right-censored. Although the exact outcome for right-censored cases is unknown, data analysis must incorporate these cases. Among other considerations, the cases with longer time periods are more likely to be censored. Following the standard for survival analysis, this study assumes that the outcome for censored cases is at least equal to, but possibly greater than, those completed by the cutoff date. The following program in a DATA step can be used to define the censorship:

```
...
cutoffdt='31mar91'd;
format cutoffdt date7.;
if tbsend=. or tbsend>cutoffdt
    then durat=tbsend-cutoffdt;
if tbsend^=. & tbsend<=cutoffdt
    then durat=tbsend-tbsstar;
censor=(durat<0);
adurat=abs(durat);
durwks=round(adurat/7,0.1);
...
```

In addition, the combined sample data contains elements that will be used as explanatory variables in multiple regression analysis. In this study, these variables include the nature of the case (i.e., treatment A or treatment B case), age, sex, injury type, years of employment, industrial location and other factors involving case processing and social demographic characteristics.

Stratified Survival Analysis

Using the combined sample data, SAS LIFETEST produced a stratified survival analysis. Treatment A and treatment B cases were classified into two strata, then subjected to the product-limit method² to compute the survival function distribution (total exit rate) against time (outcome). This stratified analysis produced two exit rate curves, one for each stratum, illustrating the probability distribution that an experimental case from the population would close at each specific time point.

The LIFETEST procedure not only provides the stratified survival function curves but also produces duration value at 25, 50 and 75 percentiles for comparison among strata. In addition, three statistical tests (log-rank, Wilcoxon and likelihood ratio) for homogeneity indicate whether the patterns of survival function show a significant statistical difference from treatment A to treatment B.

SAS LIFETEST provides a plot of estimated survival function against time. It is possible to create an original plot by default printing survival function values along with the time

² It is also called the Kaplan-Meier method; see Kalbfleisch, J.D. and Prentice, R.L. (1980), *The Statistical Analysis of Failure Time Data*, New York: John Wiley & Sons, Inc.

values. However, that could be time consuming. The following program will save time in obtaining information needed for Figure 2:

```

...
Proc lifetest plot=(s,ls,lls) outsurv=curv maxtime=116 notable;
    time durwks*censor(1);
    strata flag;                                /* A or B cases*/
data old1;
    set curv;
    where flag=1;                                /* A cases only*/
wks=round(durwks,1);
proc sort; by wks;
data old2(drop=flag durwks _censor_);
    set old1;
if wks=lag(wks) then delete;
proc print;
    id wks;
    title 'A cases: total exit rates against duration';
data new1;
    set curv;
    where flag=2;                                /* B cases only*/
...
    title 'B cases: total exit rates against time';
...

```

Run Cox Proportional Regression

The purpose of conducting a regression analysis is to ascertain that the change in outcome from treatment A to treatment B was truly because of the change in the treatment itself. By using the Cox proportional hazard model³, the following explanatory variables were included:

1. The nature of the case, classified into two categories: treatment A case, coded 0, and treatment B case, coded 1;
2. Three socio-demographic variables, including:
 - Gender (male, coded 0; female, coded 1)
 - Age (under 45, coded 0; 45 and older, coded 1)
 - Years of employment (less than 10 years, coded 0; 10 years and more, coded 1);
3. The nature of the injury, classified into three categories (fracture, sprain/strain and other);
4. The industrial allocation, classified into nine categories using the two-digit SIC code: mining, construction, manufacturing, transportation, wholesale trade, retail sales, financing and banking, services and others;

³ To verify that the proportional hazard assumption underlying the Cox model was met, the relative hazards were plotted against the duration of disability for each categorical variable. All of the included variables met the proportional hazard assumption.

5. The monthly unemployment rate, reported by Texas Employment Commission and assigned to each case according to the date case established;
6. Wage replacement, measured by the ratio of the pre-injury weekly wage replaced by weekly benefit payment. (75% or less, coded 0; more than 75%, coded 1);
7. Timeliness of payments, measured by the length of time from the onset of injury to the first benefit payment. (15 days or less, coded 0; more than 15 days, coded 1);
8. Compromise settlement: if the benefits payment was involved in a lump sum settlement or compromise settlement agreement (CSA), coded 0; otherwise, coded 1.

The general hypothesis is that if the change from treatment A to B was a major contributor to the change in outcome, by controlling for these covariates, the nature of the case (treatment A or B) should have a significant effect on the outcome.

SAS PHREG was used to conduct this regression analysis. As presented in Table 1, it provided the coefficient (β), its standard error and the relative hazard (e^{β}) of each explanatory variable. For example, the result shows that the nature of the case, a dummy variable indicating whether the case was under treatment A or B, had a significant effect on the outcome when all other explanatory variables were controlled for. The likelihood of closure under treatment B was nearly twice that under treatment A at any given time point (relative hazard=1.980).

Predicting Future Change

With the results from the Cox regression model, SAS PHREG can use a hypothetical data base to predict other changes. The following program addresses the question of what will happen if the value of significant variables changes or remains unchanged given treatment B:

```

...
data a;
...
data hypo;
    input treatment sprain payment settle;
cards;
1 0 0 0
1 1 1 1
1 1 0 0
1 1 0 1
...
;
title 'total exit rate estimates for specific covariates';
proc phreg data=a;
    model durwks*censor(1)=treatment sprain payment settle;
    baseline covariates=hypo out=pred1 survival=s;
data temp(drop=durwks);
    length scen $1;
    set pred1;

```

```

wks=round(durwks,1);
if treatment=1 & sprain=1 & payment=1 & settle=0 then scen='*'; /*reference*/
if treatment=1 & sprain=1 & payment=1 & settle=1 then scen='+';
if treatment=1 & sprain=1 & payment=0 & settle=0 then scen='#';
if treatment=1 & sprain=0 & payment=0 & settle=1 then scen='$';
...;
proc sort; by scen wks;
data final;
    set temp;
    where scen^=.;
if scen=lag(scen) & wks=lag(wks) then delete;
proc print;
    id wks;
    by scen;
proc plot;
    plot s*wks=scen;
    title 'predicted total exit rates by 4 different scenario';
...

```

Figure 3 presents the result of this predicting model. The total exit rate curve (marked by '*') represents a hypothetical scenario in which an employee suffers a sprain/strain injury, receives a late payment of benefits, and does not reach a compromise settlement. The total exit rate curves show the effect(s) of changing the indicated variable(s). For example, the curve marked by '+' represents a scenario in which an employee suffers an injury other than a sprain/strain. Other variables remain unchanged. The effect of changing that single variable reduces the likelihood that the payment will last 35 or more weeks from 48% to 40%. When all variables are changed, the cumulative effects can be substantial. The Figure 3 example assumes that the injury is not a sprain/strain, the benefits are paid in a timely fashion and a compromise settlement is made. The comparison between this curve (marked '\$') and the reference curve (marked '*') indicates that the variable changes in the later scenario reduce the probability the claim will last 35 weeks or more from 48% to 20%.

Discussion

SAS LIFETEST and PHREG procedures have proven efficient in survival analysis. Their ease of use and flexibility are solid advantages to the programs.

There are, however, at least two areas in which the programs could be improved. First, a combination of the two procedures would eliminate the researcher's need to move back and forth between them. Second, when the Cox proportional hazard model is used and the explanatory variables are of ordinal level, the computation of regression coefficients for each category of the variable needs to be provided.

References

Kalbfleisch, J. D. and R. L. Prentice. 1980. *The Statistical Analysis of Failure Time Data*. New York: John Wiley and Sons, Inc.

Meyer, B. D., W. K. Viscusi and D. L. Durbin. 1990 (October). "Workers' Compensation and Injury Duration: Evidence From a Natural Experiment." Cambridge, MA: National Bureau of Economic Research. Working paper no. 3494.

SAS Institute Inc. 1992. *SAS/STAT User's Guide*, Version 6, Fourth Edition. Vol. 2, Cary: SAS Institute Inc.

SAS Institute Inc. 1991. Technical Report P-217: *SAS/STAT Software: The PHREG Procedure*, Version 6. Cary: SAS Institute Inc.

Worrall, J. D., and D. Appel. 1982. "The Wage Replacement Rate and Benefit Utilization in Workers' Compensation Insurance," *Journal of Risk and Insurance* 49 (September): 361-371.

Figure 1: Censored and Noncensored Cases

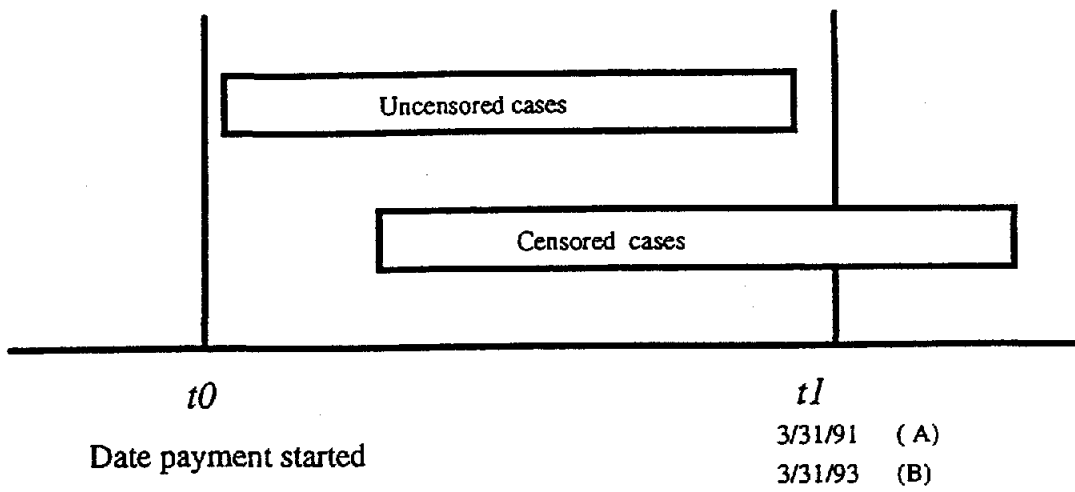


Table 1: Proportional Hazard Model: Combined Sample Data

Covariates	Coefficient (B)	Std. Error	Relative Hazard ²
Nature of case			1.00
Treatment A case (r)	0.68*	0.094	1.98
Treatment B case			
Gender			1.00
Male (r)	0.02	0.06	1.02
Female			
Age			1.00
less than 45 years	-0.04	0.059	0.96
45 years and over			
Years of employment			1.00
less than 10 years	-0.09	0.063	0.92
10 years or more			
Type of injury:			1.00
All other (r)	-0.08	0.103	0.92
Fracture	-0.15*	0.058	0.86
Sprain/Strain			
Industry:			1.00
All other (r)	-0.22	0.199	0.80
Mining/oil	-0.22	0.153	0.80
Construction	-0.20	0.145	0.82
Manufacturing			
Transportation,	-0.08	0.155	0.92
Communication	-0.13	0.167	0.87
Wholesale	-0.20	0.150	0.82
Retail sales			
Finance, Bank &	-0.35**	0.204	0.70
Real estate	-0.21	0.148	0.81
Services			
Unemployment rate:	0.02	0.056	1.02
Percent of preinjury wage replaced:			1.00
75% or less	-0.02	0.058	0.98
more than 75%			
Timeliness of payment			1.00
within 15 days after the date lost	-0.11*	0.053	0.89
time began			
more than 15 days after the date			
lost time began			
Compromise settlement			
within 15 days	0.42*	0.094	1.53
more than 15 days			

Notes:

* - significant at $p=0.01$ (two-tailed test). ** - marginal significant ($p=0.08$).

1 - (r) = the reference category.

2 - the conditional probability of ending disability relative to the reference category.

Figure 2

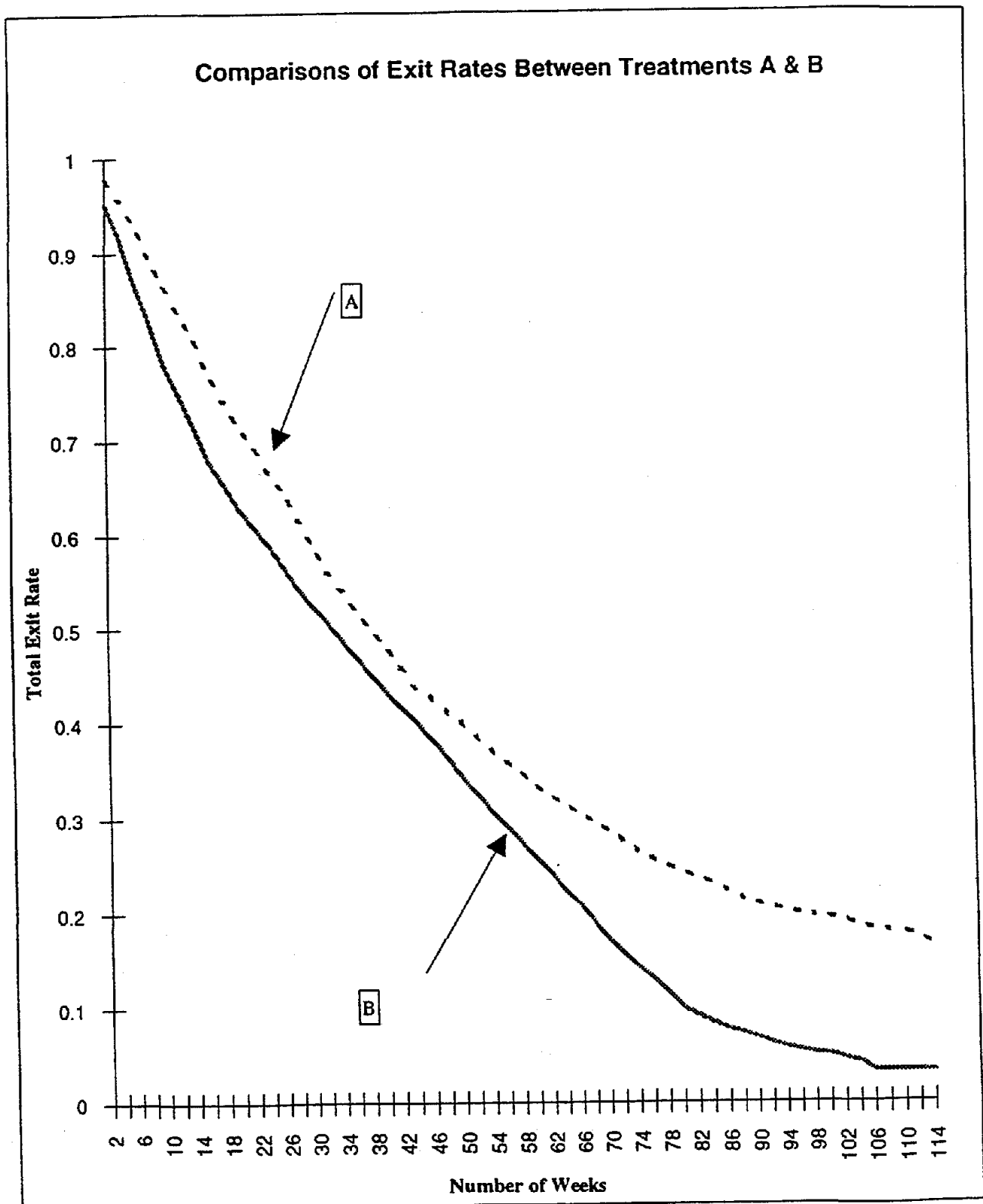
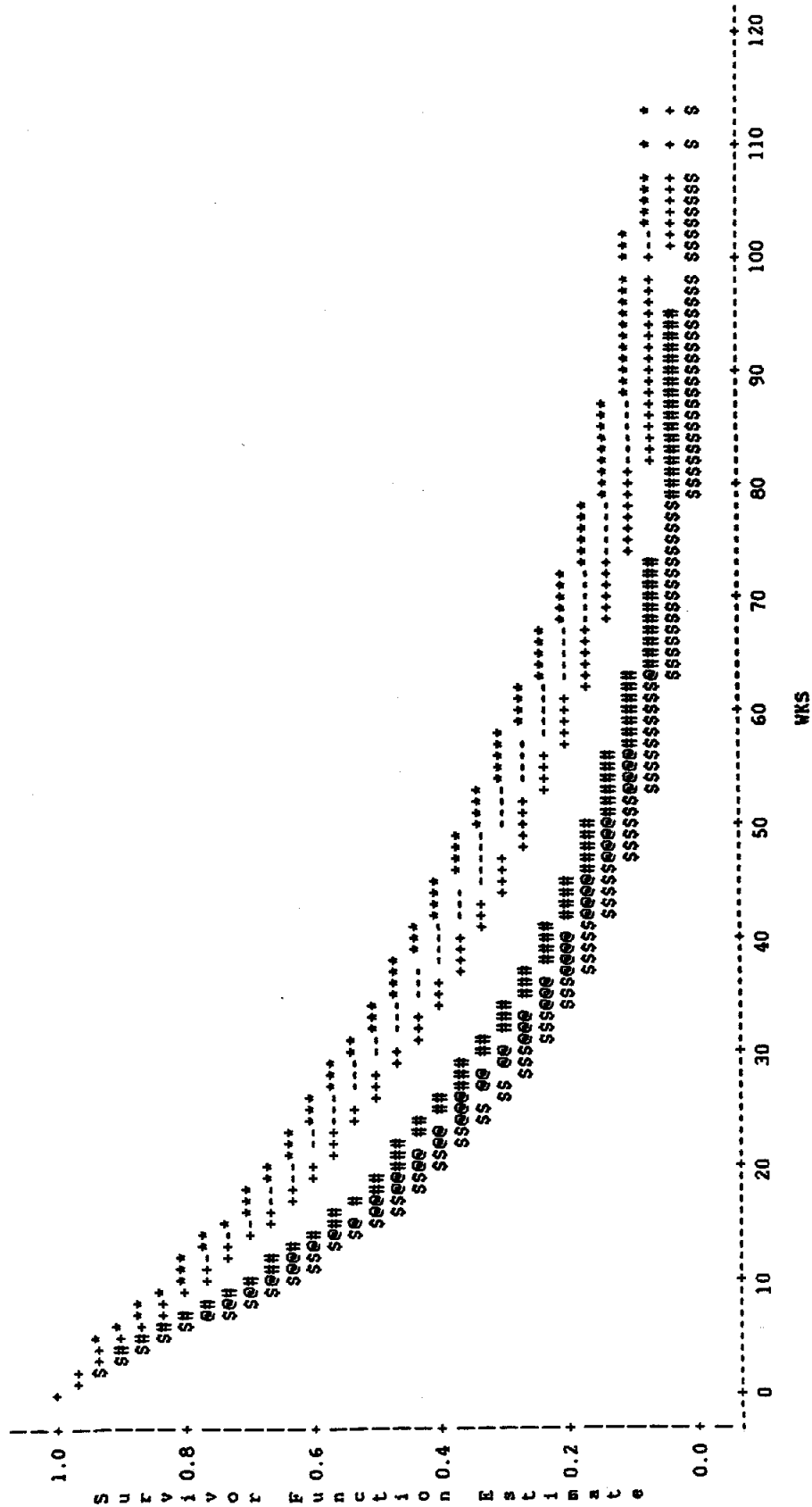


FIGURE 3. PREDICTED TOTAL EXIT RATES
BY SIX DIFFERENT SCENARIOS

Plot of S*WKS. Symbol is value of SCEN.



NOTE: 129 obs hidden.