

Bootstrapping the Kaplan-Meier Difference Equation

George W. Pasdirtz, Hazleton Laboratories

Introduction

At Hazleton Laboratories, we are looking for new ways to analyze survival data from long-term rodent carcinogenicity studies. The Kaplan-Meier (K-M) or product-limit curve (Kaplan and Meier, 1958) is a useful graphical technique for displaying survival curves, but it does not produce p-values for group differences. The other available parametric and nonparametric techniques in SAS* produce p-values but rely on assumptions that are sometimes difficult to justify scientifically. In this paper, an extension of the K-M analysis is developed which generates p-values for group differences and relies on fewer basic assumptions.

In a typical rodent carcinogenicity study, we record time to death and, for selected palpable and rapidly lethal tumors, time to tumor (Thakur, Wetzel, and Stevens, 1990) for animals given various doses of a compound. The current methods for analyzing survival data using SAS PROCs LIFEREG and LIFETEST are specially designed to handle a common feature of lifetime data, censored observations. Censoring enters either due to withdrawal of animals or termination of the study. Some animals die on test for reasons unrelated to treatment or are serially sacrificed to obtain tumor onset times. At the end of the study, the remaining animals are sacrificed. In both cases, the complete lifetimes (and thus tumor incidences) for the "censored" animals are unobserved.

There are some well known problems with conventional survival models based on maximum likelihood (ML) estimation. The parametric estimates of survival differences produced by PROC LIFEREG require distributional assumptions that a pathologist may not be in a position to verify. The choice of the extreme value, normal, logistic, exponential, Weibull, log-normal, log-logistic, or gamma distributions cannot usually be justified in applied work except that one choice might produce a better estimator on a given data set. In a production environment, there is seldom time to search for that best model. And, the standard errors produced by ML models are approximations, which can be poor in small samples. If we choose to use the nonparametric estimators in PROC LIFETEST, the arbitrariness of the underlying asymptotic theory and the expected-frequency weighting schemes make the reported p-values difficult to justify.

It should be pointed out that none of the SAS procedures use the K-M curve directly. The collection of techniques, however, are all asking the same question: Are there statistically significant differences in animal lifetimes given increasing doses of a potential carcinogen? If we could generate p-values directly from the K-M curve, with fewer appeals to extraneous assumptions, we would have greater confidence in our conclusions about a compound.

The Hazleton Data

Figure 1 shows the K-M survival curve produced by PROC LIFETEST for a 106-week rodent carcinogenicity study conducted at Hazleton Laboratories. There were five dose

groups with the zero-dose group (1) acting as the control. In a situation where there are dose-related effects on survival, the curves should be well separated and ordered from right to left on the graph, with the highest dose group having the lowest K-M survival curve.

In this particular set of data, there is no clear ordering. However, the mid-dose groups (3-4) are well separated. In some studies, we also observe that the high-dose groups will have the highest survival curve due to dietary effects (the animals lose their appetite, eat less, and have fewer diet-related tumors). Below, we will only pursue the testing of differences between experimental and control groups.

The K-M (product-limit) estimator is the cumulative product

$$S_t = \prod_{j=1}^k (1 - d_j / n_j)$$

where n_j is the number of surviving units (the number of animals at risk), and d_j is the number of animals that fail at a distinct failure times

$$t_1 < t_2 < \dots < t_k.$$

The mean survival time is given by

$$\mu = \sum_{j=1}^k S(t_{j-1})(t_j - t_{j-1})$$

where t_0 is defined to be zero. Although a standard error is computed in PROC LIFETEST from Greenwood's formula

$$\sigma(S) = S(t) \sqrt{\frac{\sum_{j=1}^k d_j}{n_j S_j}}$$

no probability estimates are attached to the K-M means. There are some questions about the appropriateness of the mean as a measure of central tendency; Greenwood's formula for the standard error, however, has withstood simulation analysis (Efron, 1981).

Another way to describe the K-M process, which will lead to a new estimator, is the demographic model

$$N_t = N_{t-1} + B_{t-1} - D_{t-1}$$

where the population at a given time point, N_t , is the difference between births and deaths, $B_{t-1} - D_{t-1}$, plus the prior population, N_{t-1} . In a typical carcinogenicity study,

where there are no births into the population and where we can distinguish withdrawals from the study by D_{i-1}^+ , we can rewrite the K-M process as

$$N_i = N_{i-1} - D_{i-1} - D_{i-1}^+$$

$$N_i^+ = N_{i-1}^+ + D_{i-1}^+$$

$$S_i = \frac{N_i}{N_0 - N_i^+}$$

a difference equation (DE) system where animals move to one of two compartments, one for withdrawals N_i^+ and one for animals that continue on test, N_i . We can substitute S_i (the number surviving as a proportion of the number of animals at risk from the K-M definition) back into the difference equation system, rearrange terms, and obtain the estimator

$$S_i = \alpha + \beta S_{i-1}$$

which has the solution (Goldberg, 1958)

$$S_k = \beta^k (S_0 - S^*) + S^*$$

$$S^* = \frac{\alpha}{1 - \beta}$$

where S^* is the equilibrium value. To generate the typical survival curve, $\alpha < 1$ and $\beta > 1$.

The SAS code for generating the difference equation model is given in Appendix 1. Data must be made available in the same form that is required for either PROCs LIFEREG or LIFETEST and must be sorted by group and by time of death. The data step computes the difference between the current observed time of death and the last time of death and then generates the appropriate compartment of the DE model.

The DE curves for each group are displayed in Figure 2. A point-by-point comparison of the DE and the K-M curves indicates that the survival values at each censored survival time are the same, except for the last time point, which is not computed in the K-M curve.

The results of estimating the model using ordinary least squares (OLS), with dummy variables included to test for group differences, are given in Appendix 2. Although the overall regression is significant, there are no significant group differences.

The naive choice of an OLS estimator might be questioned, even though it is similar in spirit to estimating mean survival time. Since the DE model is essentially a time series process, we can test for correlation among the residuals. The first order autocorrelation, 0.030, is very small. The Shapiro and Wilk (1965) statistic, W,

Moments			
Std Dev	0.003036	Variance	9.219E-6
Skewness	-2.98868	Kurtosis	14.88298
USS	0.004877	CSS	0.004877

CV	.	Std Mean	0.000132
T:Mean=0	0	Prob> T	1.0000
Sgn Rank	42161.5	Prob> S	0.0001
W:Normal	0.614988	Prob<W	0.0

however, shows significant departures from normality, as does the Normal Probability plot (Figure 3). Finally, we can run Levene's (1960) test for equality of variances, which indicates that variances from comparing groups 1 and 2 with the control are not homogeneous.

To maintain parametric assumptions, we could investigate normalizing transformations and weighted least squares (WLS) solutions to variance equalization. However, given space limitations, we will pursue a nonparametric solution that also allows us to compare the DE and ML models.

The accelerated failure time models estimated by PROC LIFEREG assume that the independent variable acts as a multiplier on the event time (death or tumor onset). In terms of survival probabilities

$$\Pr(T > t|x) = \Pr[T_0 > \exp(-x'b)t]$$

the event times are estimated in a linear model framework $-x'b$, which is comparable to the DE model.

The estimates for the accelerated failure time model are given in Appendix 3. The technique also finds no group differences. Due to space limitations, we will not explore the appropriateness of the ML model to this data set.

Bootstrap Estimates of Error and Bias

Since both approaches produce similar conclusions, we will need a better way to compare the techniques. We also need a method for circumventing normal-theory problems in the DE model. We can reach both objectives by using a nonparametric, computer-intensive technique called the bootstrap (Efron, 1979, 1982) which can be used to produce comparative estimates of bias and error.

The bias of an estimator refers to the deviation of the expected value from the true value.

$$\text{Bias}(S) = E[S] - \mu$$

The mean squared error (MSE) is the expected value of the square of the deviation of the estimator from the true value.

$$\text{MSE}(S) = \sigma^2 + [\text{Bias}(S)]^2$$

Estimators should have low bias and minimum MSE.

To compute bias and MSE, the bootstrap draws a random sample, with replacement, from the original observations (X,Y). Each model is then estimated on the new, synthetic sample, and the regression estimates are stored in a file. The process is repeated many times to develop the sampling distribution of the estimators from which the population

mean μ and variance σ^2 can be computed. The bootstrap estimators are the nonparametric ML estimates of the population values. For regression problems (Freedman,

1981), the bootstrap replicates the standard analytical solutions.

To compute the bootstrap, we use a modification of a SAS macro written by Carson (1985). The synthetic population estimates are presented in Appendix 4 and Appendix 5. Although the standard errors produced by the DE model are substantially smaller than those from the ML model, measures of bias and MSE are not scale-free and thus not really comparable. Instead, we can construct a scale-free number, the standard t -statistic

$$t = \frac{\beta}{\sigma}$$

compute measures of Bias(t), and then compute relative biases

$$\frac{\text{Bias}(t)_{DE}}{\text{Bias}(t)_{ML}}$$

The results for the intercept and four group differences,

{-0.6136, 0.32541, -3.8319, 3.9256, -0.78017}

indicate that the DE model was less biased for the comparisons 1-2 and 1-5 but more biased for the 1-3 and 1-5 comparisons, a mixed conclusion. Notice from Figure 2, that groups 2, 4, and 5 cross, a standard problem of interpretation in K-M analysis (animal populations differing only by treatment (different dose levels of the same compound) should not have survival curves of different shape). A single survival model might not be appropriate for this data set. It might also be anticipated that the appropriate WLS solution might further reduce Bias(t) for the DE model.

Varying Noncensored Survival Times

Another important comparative measure is the relative power of two tests. Power is the probability of detecting a significant difference when the null-hypothesis is actually false. A test with higher power, other things being equal, is preferable.

Power can be increased by drawing a larger sample or by decreasing the variance of the estimator. In terms of the K-M curve, increasing the sample size will increase the number of noncensored survival times and produce a curve that becomes smoother as the sample size approaches infinity.

For the K-M curve, bias and power are related since the estimator from a smooth curve will have less bias and smaller variance than an estimator derived from a series of step-functions. Preliminary simulation results indicates that as the number of noncensored survival times increases beyond 50% of the time points, bias becomes asymptotically small.

References

Carson, R. T. (1985), "SAS® Macros for Bootstrapping and Cross-Validating Regression Equations," *SUGI 10 Proceedings*, 1064-1069.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1-26.

____ (1981), "Censored Data and the Bootstrap," *JASA*, 76, 312-319.

____ (1982), *The Jackknife, the Bootstrap and Other Resampling Plans* (Philadelphia: SIAM).

Freedman, D. A. (1981), "Bootstrapping Regression Models," *The Annals of Statistics*, 9, 1218-1228.

Goldberg, S. (1958), *Introduction to Difference Equations* (New York: Wiley).

Kaplan, E. L. and P. Meier (1958), "Nonparametric Estimation from Incomplete Observations," *JASA*, 53, 457-481.

Levene, H. (1960), "Robust Tests for Equality of Variances," in *Contributions to Probability and Statistics*, (eds.) I. Olkin et. al., Ch. 25 (Stanford, CA: Stanford University Press) 278-292.

Shapiro, S. S. and M. B. Wilk (1965), "An analysis of variance test for normality (complete samples)," *Biometrika*, 52, 591-611.

Thakur, A. K., L. T. Wetzel, and J. T. Stevens (1990), "Statistical Issues in Lifetime Oncogenicity Studies in Rodents," *Proceedings of the Biopharmaceutical Section of the American Statistical Association*, 275-281.

*SAS is a registered trademark of the SAS Institute, Inc., Cary, NC, USA.

Author

George W. Pasdirtz, Ph.D.
Hazleton Laboratories
3301 Kinsman Boulevard
Madison, WI 53704

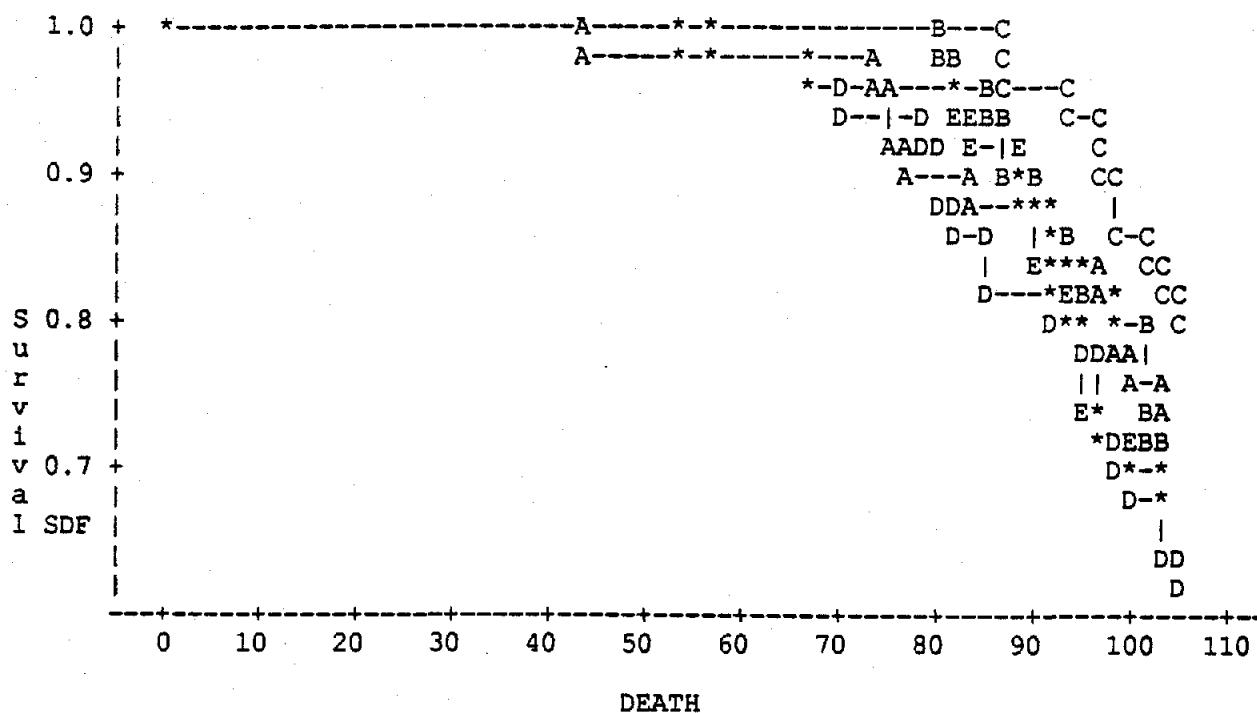


Figure 1. K-M Plot

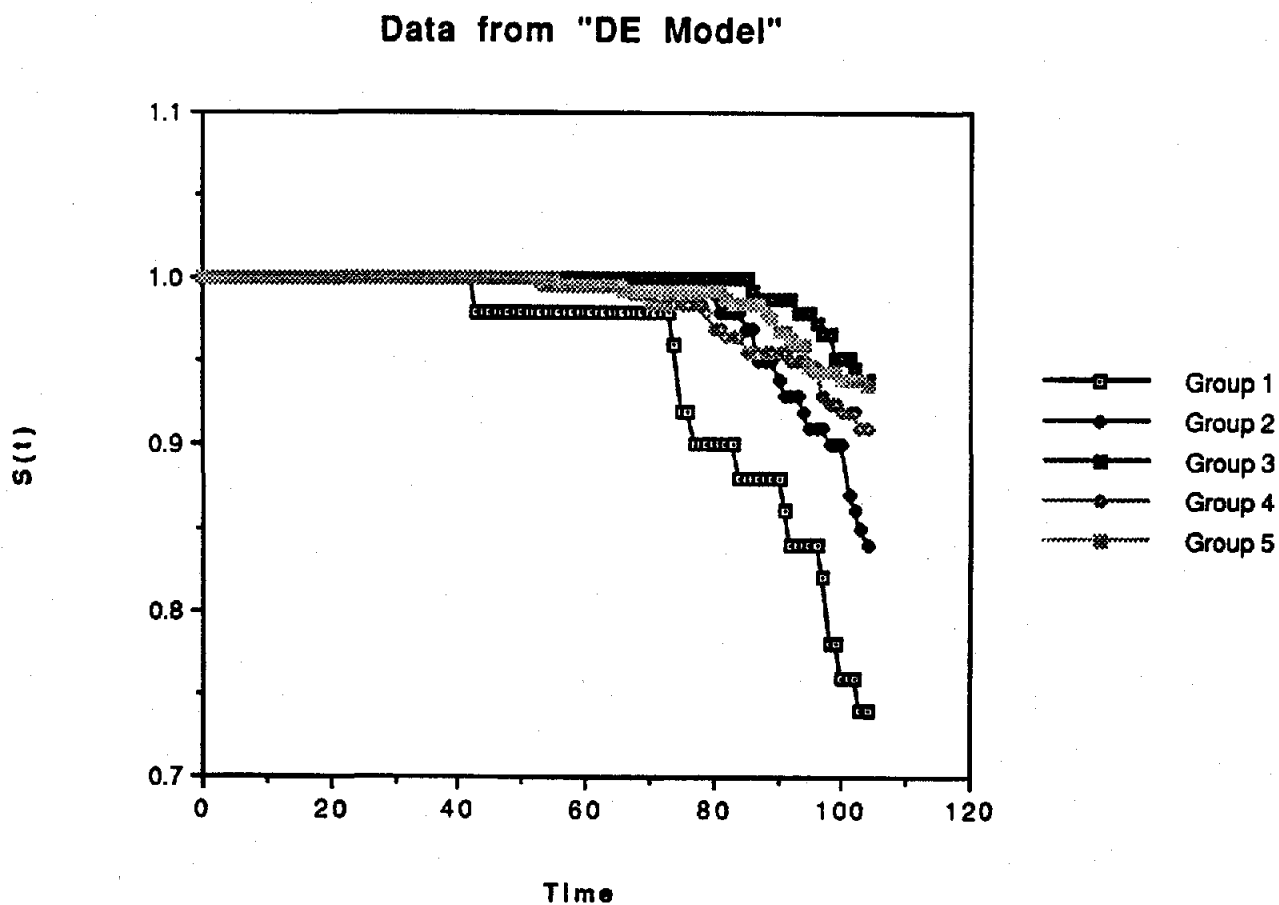


Figure 2. DE Plot

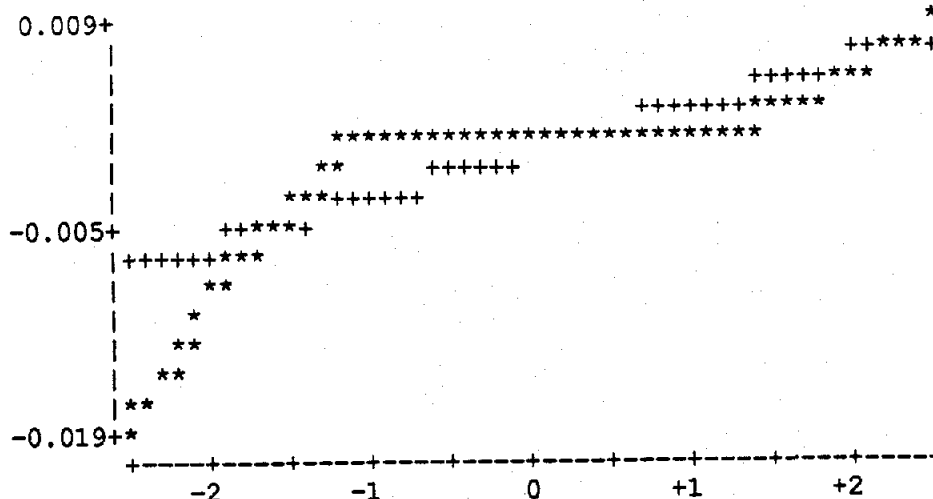


Figure 3. Normal Probability Plot

Appendix 1. SAS Code for generating the K-M difference equation.

*NOTE The data set z must be in a form similar to PROC LIFEREG or LIFETEST,
sorted by group and by time of death containing
death = time of death
censor = censored
group = grouping variable
the K-M survival curve is returned in
p = percent surviving t
pl = percent surviving t-1;

```
data x;
  set _last_ by group;
  keep n d nc dc time group p nl k n0 death censor pl;
  retain nc time d dc p n n0 pl;
  k = death-lag(death);
  if first.group then do;
    n=total; nl=total; time=0; nc=0; d=0; dc=0; k=death; n0=n; p=1; pl=1;
  end;
  nl=lag(n); ncl=0;
  do i = 1 to k;          * Interpolate to next time point;
    output; time=time+1; pl=p; * not entered if k < 1;
  end;
  if censor = 1 then do;
    nc=nc+1; dc=dc+1; n=n-1;
  end;
  else do;    n=n-1; d=d+1; end;
  nl=lag(n); pl=p; p = n / (n0-nc);
```

Appendix 2. Output from the DE Model.

Dependent Variable: P

Root MSE		0.00305	R-square	0.9957	
Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	-0.035784	0.00316107	-11.320	0.0001
PL	1	1.035654	0.00317968	325.710	0.0000
D1	1	-0.000279	0.00043835	-0.635	0.5254
D2	1	-0.000492	0.00041970	-1.172	0.2419
D3	1	-0.000186	0.00041932	-0.445	0.6568
D4	1	0.000073759	0.00041943	0.176	0.8605

Dependent Variable: YABS (Absolute Value of Residuals for Levene's Test)

Parameter	Estimate	T for H0: Parameter=0	Pr > T	Std Error of Estimate
INTERCEPT	0.0343021402	15.94	0.0001	0.00215168
PL	-.0339476463	-15.68	0.0001	0.00216435
D1	0.0010892427	3.65	0.0003	0.00029838
D2	0.0007040598	2.46	0.0140	0.00028568
D3	0.0002849043	1.00	0.3187	0.00028542
D4	-.0000215642	-0.08	0.9398	0.00028550

Appendix 3. Output from PROC LIFEREG.

L I F E R E G P R O C E D U R E

Data Set =WORK.A
 Dependent Variable=Log(DEATH)
 Censoring Variable=CENSOR
 Censoring Value(s)= 1
 Noncensored Values= 74 Right Censored Values= 175
 Left Censored Values= 0 Interval Censored Values= 0

Log Likelihood for WEIBULL -92.19560136

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	4.89320674	0.056175	7587.656	0.0	Intercept
D1	1	-0.2640418	0.164941	2.562634	0.1094	
D2	1	-0.3467961	0.214983	2.602189	0.1067	
D3	1	0.46808647	0.270016	3.005192	0.0830	
D4	1	0.78354782	0.484793	2.61227	0.1060	
SCALE	1	0.15572417	0.017289			Extreme value scale paramet

Appendix 4. LIFEREG Bootstrap.

N Obs	Variable	Label	Mean	Std Dev
101	LNLIKE	Log of Likelihood	-88.2404634	8.8471757
	INTERCEP	Intercept	4.8914496	0.0637342
	DEATH		-1.0000000	0
	D1		-0.2459542	0.1708265
	D2		-0.3206284	0.2191065
	D3		0.4357481	0.2767229
	D4		0.7186381	0.4906730

Appendix 5. DE Bootstrap.

N Obs	Variable	Label	Mean	Std Dev
101	RMSE	Root mean squared error	0.0030293	0.000243439
	INTERCEP	Intercept	-0.0358457	0.0072210
	PL		1.0357212	0.0072933
	D1		-0.000272288	0.000466187
	D2		-0.000511578	0.000292961
	D3		-0.000201045	0.000188373
	D4		0.000065475	0.000222573
	P		-1.0000000	0