TRUNCATED DATA ARISING IN WARRANTY AND FIELD PERFORMANCE STUDIES, AND SOME USEFUL STATISTICAL METHODS

J.D. Kalbfleisch and J.F. Lawless

IIQP Research Report
RR-91-02

March, 1991

TRUNCATED DATA ARISING IN WARRANTY AND FIELD PERFORMANCE STUDIES, AND SOME USEFUL STATISTICAL METHODS

J.D. Kalbfleisch and J.F. Lawless
University of Waterloo
Waterloo, Ontario, N2L 3G1

March, 1991

ABSTRACT

Truncated data arise when a variable is observable only over some portion of its range. In this note we describe how truncated data arise in studies of the field performance or reliability of manufactured items. Failure to account for truncation can lead to biased inferences. We present some useful nonparametric methods, with examples.

Key Words: lifetime data, truncated data, nonparametric estimation, field reliability, warranty data

1 Introduction

Truncated data arise when a variable is observable only if it lies in some specified portion of its range. The purpose of this note is to show how truncated data can occur in studies of the field performance or reliability of manufactured items. Failure to account for truncation can lead to biased inferences. A second purpose is to present some useful nonparametric methods that have been developed in biomedical contexts and are useful for the problems we discuss. The paper complements the recent article by Nelson (1990). We begin with an example to motivate what follows.

Example 1. Right-truncated field reliability data

Manufactured items are sold and enter service over time. Field reliability data are often collected in such a way that only items satisfying certain conditions are observed. Kalbfleisch and Lawless (1988) describe a situation where time to first failure (referred to as the failure time) and covariate information are collected only for those items that fail over a given calendar time period $(0, T)$. In this situation, an item that enters service at calendar time u_i in $(0, T)$ and has failure time Y_i is observable if and only if $Y_i \leq \tau_i$, where $\tau_i = T - u_i$. This provides one example of right-truncation. It is key that we find out about an item only when it fails, as for some warranty or failure reporting schemes. If, on the other hand, we know about all items that enter service then an item's failure time is merely censored if $Y_i > \tau_i$. The difference between censoring and truncation is discussed further in Section 5.

To specify truncation more formally, suppose that observations are taken on a continuous random variable Y which has probability density function (p.d.f.) $f(y)$, cumulative distribution function (c.d.f.) $F(y) = Pr(Y \leq y)$ and survivor function $\bar{F}(y) = Pr(Y \geq y)$.

A *left-truncation* mechanism is said to operate when there is associated with Y a truncation variable τ such that Y is observable only if $Y \geq \tau$. Independent observations arise as pairs (Y_i, τ_i), and conditional on τ_i, the distribution of Y_i has density $f(y_i)/\bar{F}(\tau_i)$ with $y_i \geq \tau_i$. The likelihood function is

$$L = \prod_{i=1}^{n} \frac{f(y_i)}{\bar{F}(\tau_i)} \tag{1}$$

for a random sample of n observations. Similarly, *right-truncated* data occur in pairs (Y_i, τ_i) with $Y_i \leq \tau_i$ and give the likelihood function

$$L = \prod_{i=1}^{n} \frac{f(y_i)}{F(\tau_i)} \; . \tag{2}$$

Simultaneous left- and right-truncation is also possible, and is defined in the obvious way.

We now describe two specific instances of truncated data arising with field reliability and warranty studies.

Example 2. Data on the brake pad life of automobiles

In a field study to estimate brake pad life for a particular car line, the manufacturer used its dealer network to select cars at random. The remaining brake pad thickness (actually the maximum thickness over a specified set of locations on the front brake pads) was measured for each car. Let $W \geq 0$ represent the total brake wear at the time of sampling; $W = 0$ represents no wear and $W = 1$ represents the level of brake wear that requires replacement of the pads. One can think of W values greater than 1 as accumulating additional wear on subsequent pads. We assume that for a given car W depends linearly with negligible error on the accumulated mileage and, given the mileage, W does not depend on time in service. Under these assumptions, an imputed brake pad lifetime $Y = \tau/W$ can be computed for each car sampled, where τ is the mileage accrued at the time of sampling. Of particular interest is the estimation of the distribution of Y, or equivalently, of the slope of the line which expresses the mileage τ as a function of the wear W.

Estimation of the distribution of W is complicated by the following: if a selected car had already worn out a set of brake pads, it was discarded from the sample. Thus, cars with rapidly wearing brake pads are underrepresented in the sample. It follows that W is sampled if and only if $W < 1$ (right-truncation at 1). If W (or equivalently τ) is distributed independently of the slope Y, then this translates into a left-truncation of the distribution of Y at the current mileages τ_i for the sampled cars. Then, for the ith car, $Y_i \geq \tau_i$ and the data Y_i are left-truncated. Figure 1 illustrates the assumptions being made here.

More formally, observations y are subject to the constraint $\tau/y = w < 1$. Thus, if $f(y,\tau)$ is the joint p.d.f. of Y and τ, the conditional density of y given τ and $\tau/y < 1$ is

$$f(y|\tau, \frac{\tau}{y} < 1) = f(y,\tau) / \int_{\tau}^{\infty} f(y,\tau)dy \ .$$

If $f(y,\tau) = f_1(y)f_2(\tau)$, this reduces to the left-truncated density

$$f_1(y) / \int_{\tau}^{\infty} f_1(y)dy = f_1(y)/\bar{F}_1(\tau) \ .$$

This example is discussed further in Section 3.

Example 3. Reporting delays in warranty data

Manufacturers maintain data bases in which warranty claims and other information are recorded. A problem that hampers the timely presentation of information is the presence of a *reporting delay*, namely the time between event occurrence (e.g. generation of a warranty claim) and the recording of the event in the data base. In many warranty record systems, delays of three months or longer are not uncommon. As a result, the number of events reported as occurring in recent time intervals is lower than it should be.

One approach to this problem is to report data on a delayed basis. For example, at mid-month we might report warranty claims made up to the end of the second last month so that only claims with reporting delays exceeding 1.5 months would be missed. Another approach, aimed at presenting data in as timely a way as possible, is to adjust the recent data upwards to account for the reporting delays. Suppose that the reporting delay for the i'th claim is Y_i and that the Y_i's are independent and identically distributed with probability function $f(y)$ and c.d.f. $F(y)$; we assume for simplicity that y is measured in days ($y = 0, 1, 2, \ldots$). Let T represent the current time and $N(t; T)$ be the number of claims that have been reported to occur on day $t \leq T$. A natural estimate of $N(t) = N(t; \infty)$, the number of claims actually occurring on day t (and eventually to be reported) is

$$\widehat{N}(t) = \frac{N(t;T)}{F(T-t)} \, . \tag{3}$$

This is motivated on the grounds that only the fraction $F(T-t)$ of claims with reporting delays $y_i \leq T - t$ have been reported by day t. See Kalbfleisch, Lawless and Robinson (1991) for more detail and a treatment that deals with the ages of the units on which claims are made.

In practice, we need to estimate the reporting delay probabilities $F(y)$ for the system; this leads to a truncated data problem. Suppose that n claims have occurred and been reported up to day T. If the claims occurred on days $t_1, \ldots, t_n$ and were reported on days $t_i + y_i$ ($i = 1, \ldots, n$) then y_i must satisfy $y_i \leq \tau_i$, where $\tau_i = T - t_i$. Thus, y_i is an observation from the right-truncated distribution $f(y)/F(\tau_i)$, $0 \leq y \leq \tau_i$.

The remainder of the paper discusses ways of handling truncated data. Section 2 presents nonparametric estimates of probability distributions and ways of checking parametric models. Section 3 discusses estimation for Example 2 and 3 above. Section 4 concludes with some comments on the limitations of truncated data and other points. Truncated data have also been discussed recently by Nelson (1990), who gives a very careful discription of hazard plots for left-truncated data. The current paper deals more broadly with truncation in field reliability studies and includes a variety of techniques and plots.

2 Estimation of distributions from truncated data

Consider a random sample (y_i, τ_i), $i = 1, \ldots, n$ where the measurement y_i of interest is right-truncated at τ_i (i.e. $y_i \leq \tau_i$). If a parametric model $f(y;\theta)$ for the untruncated density of Y is used, then θ and f may by estimated by maximizing the likelihood function $L(\theta)$ given by (2). Maximum likelihood and other estimation techniques have been widely discussed for models such as the normal, exponential and gamma distributions (e.g. see Johnson and Kotz 1970, Sections 13.7, 14.7, 15.7, 17.8, Schneider 1986 or Kulldorf 1961, Chapter 3) and are in principle straightforward; see Nelson (1990) for some additional references and references to statistical software.

Our main objective here is to describe simple nonparametric estimates and ways of checking parametric models and independence assumptions. If $\tau = \max(\tau_i)$, then the best we can actually do nonparametrically is to estimate

$$G(y;\tau) = \frac{F(y)}{F(\tau)}\,, \qquad y \leq \tau\,.$$

A nonparametric estimate of $G(y;\tau)$ analogous to the empirical c.d.f. for ordinary (untruncated) data was developed by Lynden-Bell (1971) and subsequently studied by Woodroofe (1985), Wang, Jewell and Tsai (1986), Lagakos, Barraj and De Gruttola (1988), Keiding and Gill (1990), Kalbfleisch and Lawless (1991) and others. Let $y_1^*, \ldots, y_m^*$ denote the distinct y-values among $y_1, \ldots, y_n$; then the nonparametric estimate is

$$\widehat{G}(y;\tau) = \prod_{j:y_j^*>y} \left(1 - \frac{d_j}{n_j}\right) \tag{4}$$

where d_j is the number of y_i's equal to y_j^* and n_j is the number of (y_i, τ_i) pairs satisfying $y_i \leq y_j^* \leq \tau_i$. This estimator is obtained for discrete models by noting that

$$G(y;\tau) = \prod_{j=y+1}^{\tau} [1 - g(j)]$$

where $g(y) = f(y)/F(y) = P\{Y = y | Y \leq y\}$, $y = 0, 1, 2, \ldots$, and estimating $g(j)$ with d_j/n_j. An asymptotic variance estimate for $\widehat{G}(y;\tau)$ is given by

$$\widehat{\mathrm{var}}\left\{\widehat{G}(y;\tau)\right\} = \widehat{G}(y;\tau)^2 \sum_{j:y_j^*>y} \frac{d_j I(n_j > d_j)}{n_j(n_j - d_j)} \tag{5}$$

where $I(n_j - d_j)$ equals 1 if $n_j > d_j$ and 0 if $n_j = d_j$. The similarity with the product-limit estimator of the survivor function should be noted (cf. Cox and Oakes, 1984, Section 4.2).

For left-truncated data (y_i, τ_i) with $y_i \geq \tau_i$ $(i = 1, \ldots, n)$ parametric maximum likelihood is based on (1). Nonparametrically all that can be estimated is the truncated survivor function $\bar{G}(y;\tau^*)$ where $\tau^* = \min(\tau_i)$ and

$$\bar{G}(y;\tau^*) = \frac{\bar{F}(y)}{\bar{F}(\tau^*)}\,, \qquad y \geq \tau^*\,.$$

If $y_1^*, \ldots, y_m^*$ are the distinct values among $y_1, \ldots, y_n$ then the nonparametric estimate is

$$\hat{\bar{G}}(y;\tau^*) = \prod_{j:y_j^*<y} \left(1 - \frac{d_j}{n_j}\right) \tag{6}$$

where d_j is the number of y_i's equal to y_j^* and n_j is the number of (y_i, τ_i) pairs satisfying $\tau_i \leq y_j^* \leq y_i$. The estimate (6) is in fact an extension of the product-limit estimator (cf. Cox and Oakes, 1984, page 178). Nelson (1990) considers an alternative empirical hazard function estimator. In addition, (4) and (6) are equivalent since, if observations y_i are left-truncated at τ_i, then equivalent observations $y_i' = w(y_i)$, where $w(\cdot)$ is a strictly decreasing function, are right-truncated at $\tau_i' = w(\tau_i)$. A variance estimate for $\hat{\bar{G}}(y;\tau)$ is given by the right hand side of (5) with $\widehat{G}(y;\tau)$ replaced by $\hat{\bar{G}}(y;\tau^*)$ and the sum running over $j : y_j^* < y$.

The nonparametric estimates portray what is actually estimable from the data when independence is all that is assumed. In many instances, a parametric model $f(y;\theta)$ is of interest. There are two simple ways to check the hypothesized model. One is to compare plots of $G(y;\tau,\hat{\theta}) = F(y;\hat{\theta})/F(\tau;\hat{\theta})$ and $\widehat{G}(y;\tau)$ for right-truncated data and to compare $\bar{G}(y;\tau^*,\hat{\theta})$ and $\hat{\bar{G}}(y;\tau^*)$ for left-truncated data. The other is to examine residuals based on the fitted parametric models. For right-truncated data, consider

$$e_i = \frac{F(y_i;\hat{\theta})}{F(\tau_i;\hat{\theta})}, \qquad i = 1,\ldots,n \tag{7}$$

which should look roughly like a random sample from the uniform distribution on (0,1). For left-truncated data the e_i's are defined with $\bar{F}$ replacing F.

3 Examples

We now use the tools discussed in the preceding section to deal with the problems introduced in Examples 2 and 3.

Example 2 continued

The plan outlined earlier was implemented to obtain data on 98 cars; coded values for the corresponding odometer readings in kilometers, τ_i and imputed lifetimes, y_i (with $y_i \geq \tau_i$ by design) are reported in Table 1. Since $\tau^* = \min(\tau_i) = 6951$, the nonparametric estimator (6) estimates $\bar{F}(y)/\bar{F}(6951) = P\{Y \geq y | Y \geq 6951\}$ under the assumption that Y and τ are independent as discussed earlier. If defective brake pads are rare, it is reasonable to suppose that $\bar{F}(6951) = 1$ and the nonparametric estimate (6) is shown in Figure 2 under this assumption.

We note from the nonparametric estimate that the right tail of the distribution of pad life is longer than the left. This suggests that we consider parametric models such as the Weibull or lognormal. For example, a lognormal model, $\log Y_i \sim N(\mu, \sigma^2)$ yields maximum likelihood estimates from the likelihood (1) of $\hat{\mu} = 11.00$ and $\hat{\sigma} = .4368$. The estimated survivor function $\hat{\bar{F}}(y) = 1 - \Phi\left[(\log y - \hat{\mu})/\hat{\sigma}\right]$, where $\Phi(z)$ is the standard normal c.d.f., is shown as the smooth curve in Figure 2. The agreement with the nonparametric estimate is good. We remark that a Weibull distribution provides a somewhat poorer fit.

An alternative plot when $\tau^* = 0$ is to give a probability plot or hazard plot (Nelson 1990); the former amounts to plotting estimated c.d.f's on log normal probability paper. This approach cannot be used if $\tau^* > 0$, but residual plots are still available. A further check on the lognormal model is illustrated in Figure 3, where the ordered values of the uniform residuals,

$$e_i = \frac{\bar{F}(y_i; \hat{\mu}, \hat{\sigma})}{\bar{F}(\tau_i; \hat{\mu}, \hat{\sigma})}, \tag{8}$$

analogous to (7) are plotted against the uniform quantiles $(i - .5)/98$, $i = 1, \ldots, 98$. The plotted points lie close to a line with unit slope.

Another point of interest concerns the assumption that Y_i's distribution is independent of τ_i, except for the fact that it is left-truncated at τ_i. In some situations this might be questionable: for example, if recalled cars had been sold around the same time then cars with larger τ_i values might have experienced more long distance driving, possibly giving less wear on the brakes. Figure 4 shows a plot of the e_i residual (8) vs. $\log \tau_i$ for the 98 cars. There is nothing to suggest a connection of the sort mentioned.

The lognormal model shown in Figure 2 gives estimated 10th, 50th and 90th percentiles of brake pad life of approximately 34, 60 and 105 thousand km, respectively. Estimates based on the nonparametric approach are 32, 61 and 101.5 km, which are in good agreement. Confidence limits can be obtained for either the parametric or nonparametric cases but we will not pursue this here. See Lawless (1982, pages 233 ff, 411) for details on how to proceed. We do note that the agreement between the two c.d.f.'s in Figure 2 is well within sampling variation.

Finally, we remark that if the left-truncation were (improperly) ignored here, then brake pad lifetime using either the parametric or nonparametric approach would be slightly overestimated. For example, nonparametric estimates of the 10th, 50th and 90th percentiles would be based on the ordinary empirical c.d.f. for the y_i's; these are approximately 30, 64.5 and 103. The degree of overestimation in this case turns out to be insignificant relative to the differences between the parametric and nonparametric approaches, and to the precision of either approach as reflected by confidence limits; this is a consequence of the fact that most truncation times τ_i in the data were not small relative to typical brake pad lives.

Example 3 continued

Warranty claims on a single system of one car model were recorded. In total, 36,683 cars were sold, giving rise to 5,760 claims. This example is also considered in Kalbfleisch, Lawless and Robinson (1991) where a more detailed analysis that addresses problems in estimation of the claims distribution is considered. For the present analysis, we concentrate on the estimation of the distribution of reporting delays, questions of stationarity of the distributions, and adjusted estimates of the total number of claims.

For these analyses, time 0 is taken to be the day at which the first claim occurs. Figure 5 shows estimates of the reporting delay distribution based on the data available at $T =$ 91, 182, 273, 365 and 456 days. The estimates are based on (4) with $\tau = T$ and provide estimates of the c.d.f. of the delay distribution under the assumption, in each case, that $F(T) = 1$. These plots exhibit no systematic difference in the reporting delay distribution over chronological time. We remark that the roughness of the estimate for $T = 91$ near its right-hand end is because for j close to T the number of cars n_j (see (4)) observed for at least j days is small. The variance of the estimate is also large here and no difference in reporting delays is indicated.

To illustrate the use of the reporting delay probabilities, Figure 6 gives adjusted estimates using (3), of the cumulative number of claims actually made up to day t ($0 \leq t \leq 182$) along with a plot of the observed number reported by day 182. The reporting delay estimates are based on the data that were available at day 182. The observed numbers for t near 182 are substantially less than the adjusted estimates; the graph of the actual total number of claims eventually reported is very close to the latter.

4 Additional Remarks

Truncated data arise in many application areas, and if truncation is not properly accounted for, estimates can be inaccurate and misleading. It is important to ask whether there are conditions which the response variable must satisfy before an item is included in the data set. This identifies the truncation in both of Examples 2 and 3 which are discussed here, and in other application areas.

Truncation and censoring are sometimes confused. As noted above, truncation arises when the response variable for an item must satisfy certain conditions before the item is observed at all. Censoring, on the other hand, arises when the specific response cannot be observed; in this case the item is observed, but we know only that the actual value of the response falls in some interval. Suppose, for example, that units are observed until failure but that observation begins 1000 hours after the unit is first put into service. If the total number of units in service at time 0 is unknown and only units surviving at 1000 hours are observed, this is left-truncation of the failure time data. If on the other hand, it is known that N items were initially put in service, then any item not observed to be surviving at 1000 hours must have failed at some time earlier. These items are left-censored at 1000 hours. If $f(y;\theta)$ is the p.d.f. of the time to failure, the likelihood in the former case is

$$\prod_{i=1}^{n}\left[f(y_i;\theta)/\bar{F}(1000;\theta)\right] \tag{9}$$

where $y_1,\ldots,y_n$ are the observed failure times ($y_i > 1000$) for the n items observed. In the latter case, the likelihood is

$$\prod_{i=1}^{n} f(y_i;\theta)\left[\bar{F}(1000;\theta)\right]^{N-n} \tag{10}$$

The information on θ is often much greater based on (10) than on (9). Kalbfleisch and Lawless (1988) provide some examples. Nelson (1990) also stresses the distinction between truncation and censoring.

The methods discussed above for truncated data can be extended to fit regression models for the respose Y given covariates $\boldsymbol{x}$. For example, parametric regression models can be fitted to left-truncated data by using the likelihood

$$L(\theta) = \prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i;\theta)/F(\tau_i|\boldsymbol{x}_i;\theta) \tag{11}$$

where $f(y|\boldsymbol{x})$ and $\bar{F}(y|x)$ are the density and survivor functions of Y given $\boldsymbol{x}$. The principal difficulties that arise are computational; most standard software does not fit truncated likelihoods like (11). Relatively simple analyses for data subject to left-truncation can be based on the proportional hazards model as discussed by several authors (e.g. Cox and Oakes, 1984, page 178 and Lawless, 1982, page 344). Simply by reversing the time scale a "reverse time proportional hazards model" can similarly be applied to data which are subject only to right-truncation and Kalbfleisch and Lawless (1991) give some discussion and examples in a medical context.

Data that are subject to both left- and right-truncation cannot be handled by the methods in this paper. Turnbull (1976) presents an algorithm for nonparametric estimation in the univariate case but more work is needed to provide convenient calculation of confidence limits and regression methodology.

Finally, in both Examples 2 and 3, the basic model is bivariate and it is assumed that a type of independence is present to reduce the problem to one of truncation. In Example 2, for instance, the left-truncation for the distribution of the imputed brake pad lifetime Y_i arises from the assumption that the Y_i and the current mileage τ_i are independent. A similar independence between the process generating the warranty claims and the reporting delay is needed in Example 3. Informal checks for this independence can be based on residual plots as outlined in Section 3. More formal tests could be derived using regression models, but further work is needed to allow regression modelling of, for example Y_i on τ_i with simple interpretation of the results.

Acknowledgements

This research was supported in part by grants to both authors from the Manufacturing Research Corporation of Ontario and the Natural Sciences and Engineering Research Council of Canada. They thank Mike Brajac and Jeff Robinson for helpful discussions and Carol Leppinen for computational assistance.

References

Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data*, New York: Chapman and Hall.

Johnson, N.L. and Kotz, S. (1970). *Continuous Univariate Distributions, Vol. I*, New York: John Wiley.

Kalbfleisch, J.D. and Lawless, J.F. (1988). Estimation of reliability from field performance studies (with discussion), *Technometrics*, **30**, 365–388.

Kalbfleisch, J.D. and Lawless, J.F. (1991). Regression models for right truncated data with application to AIDS incubation times and reporting lags. *Statistica Sinica* **1**, 19-32.

Kalbfleisch, J.D. and Lawless, J.F. and Robinson, J.A. (1991). Methods for the analysis and prediction of warranty claims. *Technometrics*, **33**, to appear.

Keiding, N. and Gill, R. (1990). Random truncation models and Markov processes. *Ann. Statist.*, **18**, 582–602.

Kulldorff, G. (1961). *Estimation From Grouped and Partially Grouped Samples*, New York: John Wiley.

Lagakos, S., Barraj, L.M. and De Gruttola, V. (1988). Nonparametric analysis of truncated survival data, with applications to AIDS. *Biometrika*, **75**, 515–523.

Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*, New York: John Wiley.

Lynden-Bell, D. (1971). A method of allowing for known observational selection in small samples applied to 3CR quasars. *Monthly Notices of the Royal Astronomical Society*, **155**, 95–118.

Nelson, W. (1990). Hazard plotting of left truncated life data. *J. Quality Tech.*, **22**, 230–238.

Schneider, H. (1986). *Truncated and Censored Samples from Normal Distributions*, New York: Marcel Dekker.

Turnbull, B. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. Roy. Statist. Soc. B*, **38**, 290–295.

Wang, M.C., Jewell, N.P. and Tsai, W.Y. (1986). Asymptotic properties of the product limit estimate under random truncation, *Ann. of Statist.*, **14**, 1597–1605.

Woodroofe, M. (1985). Estimating a distribution function with truncated data, *Ann. Statist.*, **13**, 163–177.

Table 1. Imputed Brake Pad Life (y) and Odometer Readings (τ) for 98 cars

y (km)	τ (km)	y (km)	τ (km)	y (km)	τ (km)
22,207	38,701	18,264	24,818	30,863	42,415
23,002	49,173	17,694	68,762	22,350	34,346
23,982	42,409	20,014	68,762	44,976	106,569
28,551	73,823	13,152	89,100	18,169	20,758
21,789	46,738	16,886	64,979	30,164	52,003
17,042	44,071	14,894	65,127	21,822	77,179
25,997	61,904	15,531	59,289	18,201	68,934
23,220	39,327	6,951	53,926	22,895	78,661
18,854	49,828	15,841	79,370	27,189	165,543
21,857	46,314	14,974	47,385	10,915	79,547
27,321	56,150	38,292	61,395	25,503	55,009
13,767	50,549	11,204	72,826	12,350	46,774
23,982	54,930	38,156	53,980	39,869	124,526
20,110	54,039	26,652	37,220	17,693	92,504
15,749	49,170	17,101	44,224	26,296	109,986
26,821	44,795	28,953	50,826	14,091	101,161
27,934	72,238	18,325	65,460	21,011	59,422
15,292	107,783	18,391	86,726	11,201	27,772
28,843	81,609	18,220	43,819	10,757	33,598
15,985	45,228	15,896	100,605	25,692	69,038
23,580	124,637	16,447	67,615	32,372	75,222
53,770	64,018	23,642	89,542	13,592	58,373
21,731	82,957	19,170	60,266	19,102	105,610
28,844	143,550	23,257	103,580	16,112	56,158
17,046	43,382	20,428	82,570	53,281	55,913
16,506	69,644	20,947	87,960	57,298	83,770
15,696	74,750	28,462	42,385	36,450	123,468
27,959	32,881	23,210	68,914	19,651	68,994
13,272	51,483	17,900	95,666	20,755	101,869
16,482	31,767	46,134	78,135	30,788	87,627
24,210	77,633	39,300	83,643	20,000	38,790
17,626	63,745	11,768	18,617	39,620	74,734
27,770	82,965	17,717	92,629		

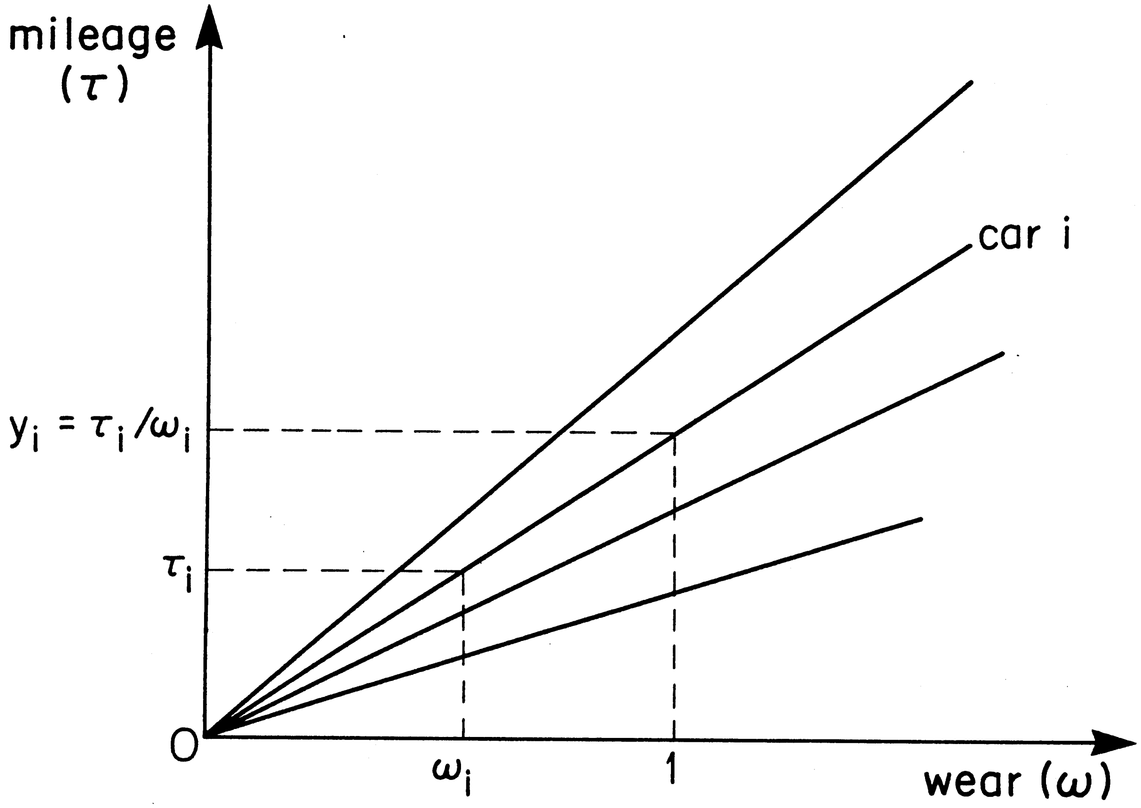


Figure 1. The linear relationship between wear w and mileage τ is illustrated. The data point (w_i, τ_i) yields the imputed lifetime $y_i = \tau_i/w_i$. The truncation of w at 1 (i.e. $w \leq 1$) translates into a left-truncation of y at $\tau(y \geq \tau)$ under the assumption that τ, y are independent. See Example 2.

Figure 2. Parametric and nonparametric estimates of the survivor function $\bar{F}(y)$ of brake pad life.

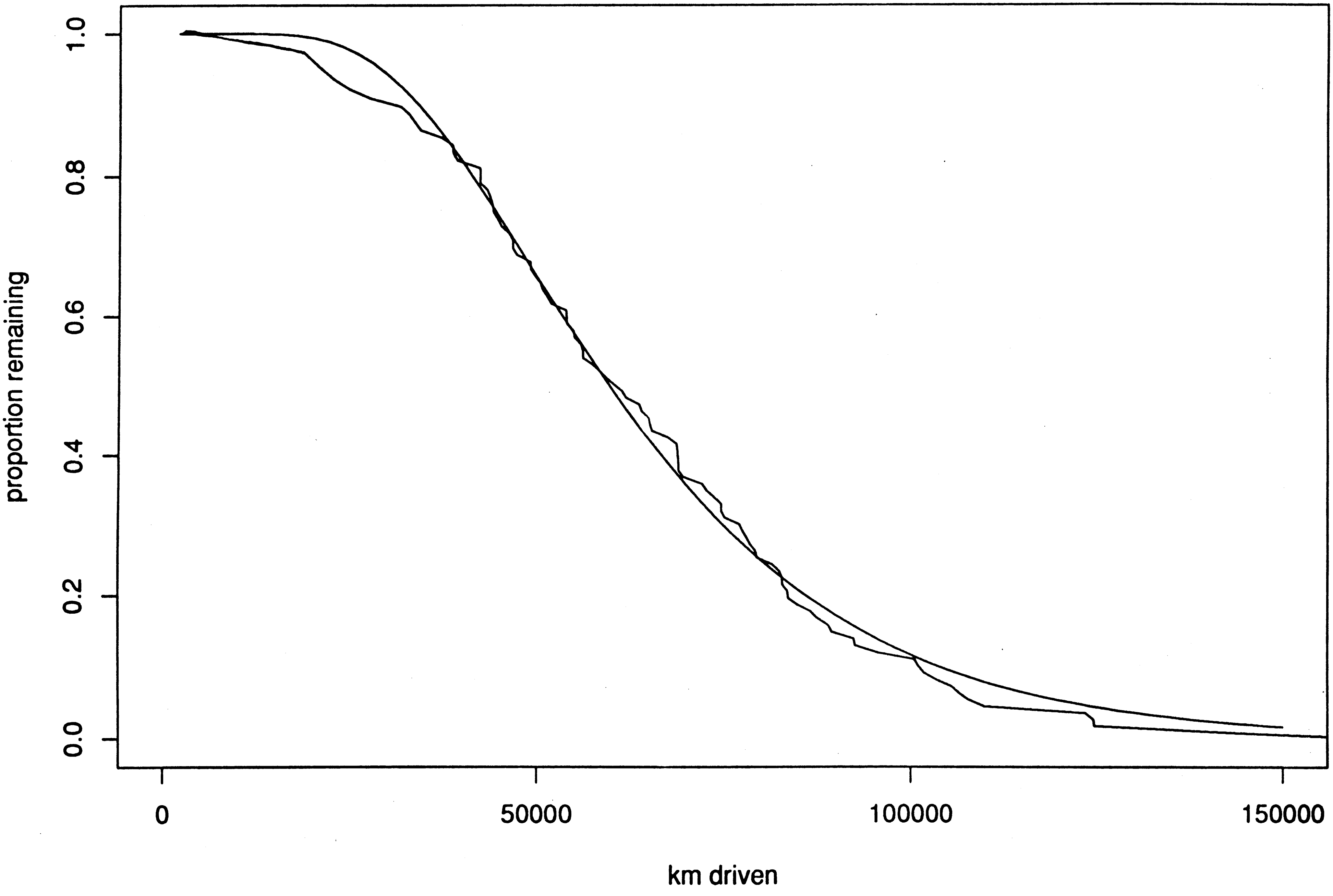


Figure 3. Probability plot of residuals (8) for the fitted lognormal distribution of brake pad life.

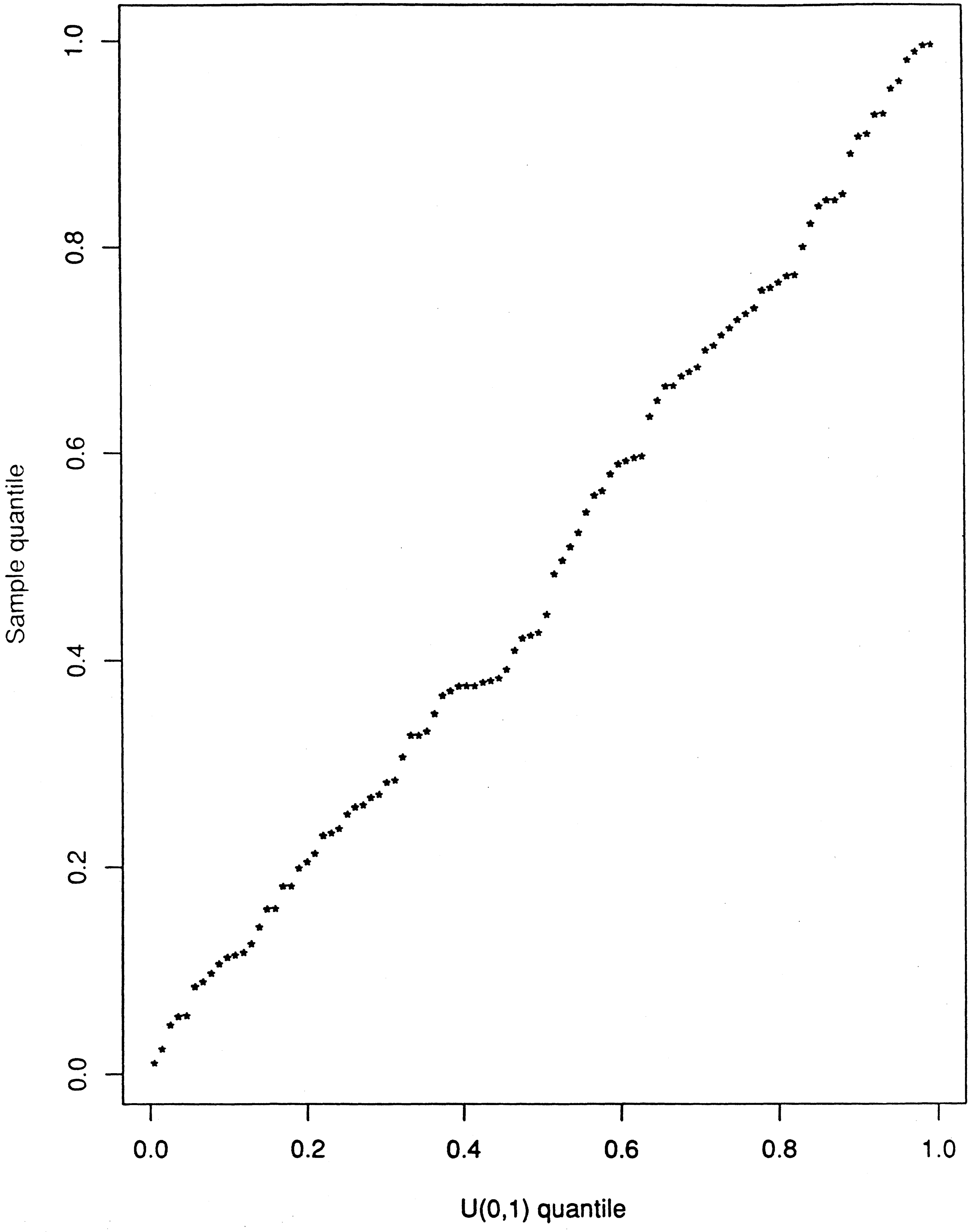


Figure 4. Plot of residuals (8) versus log truncation mileage ($\log \tau_i$).

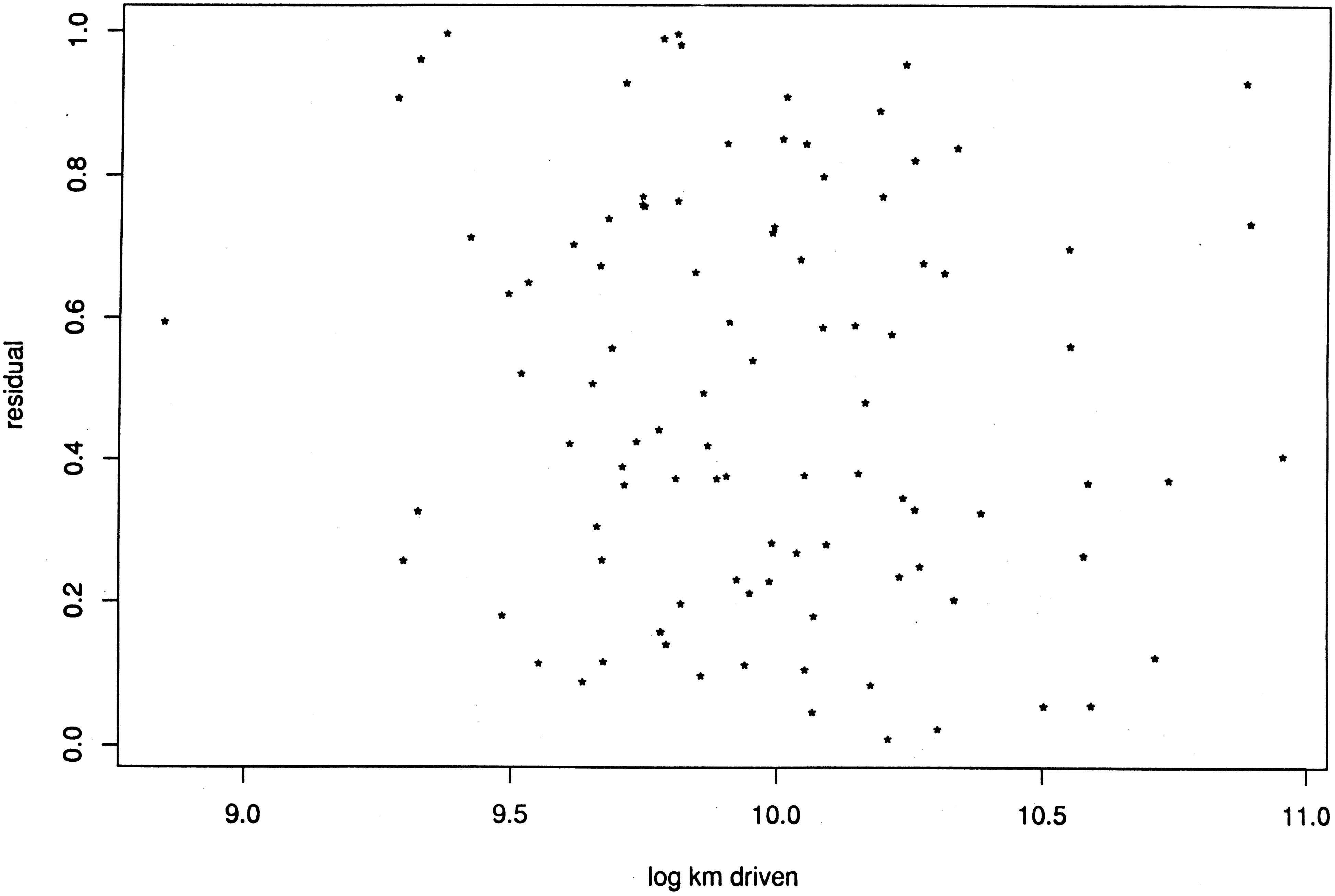


Figure 5. Estimated reporting delay cumulative distribution functions based on claims reported up to day T, for $T = 91, 182, 273, 365$ and 456.

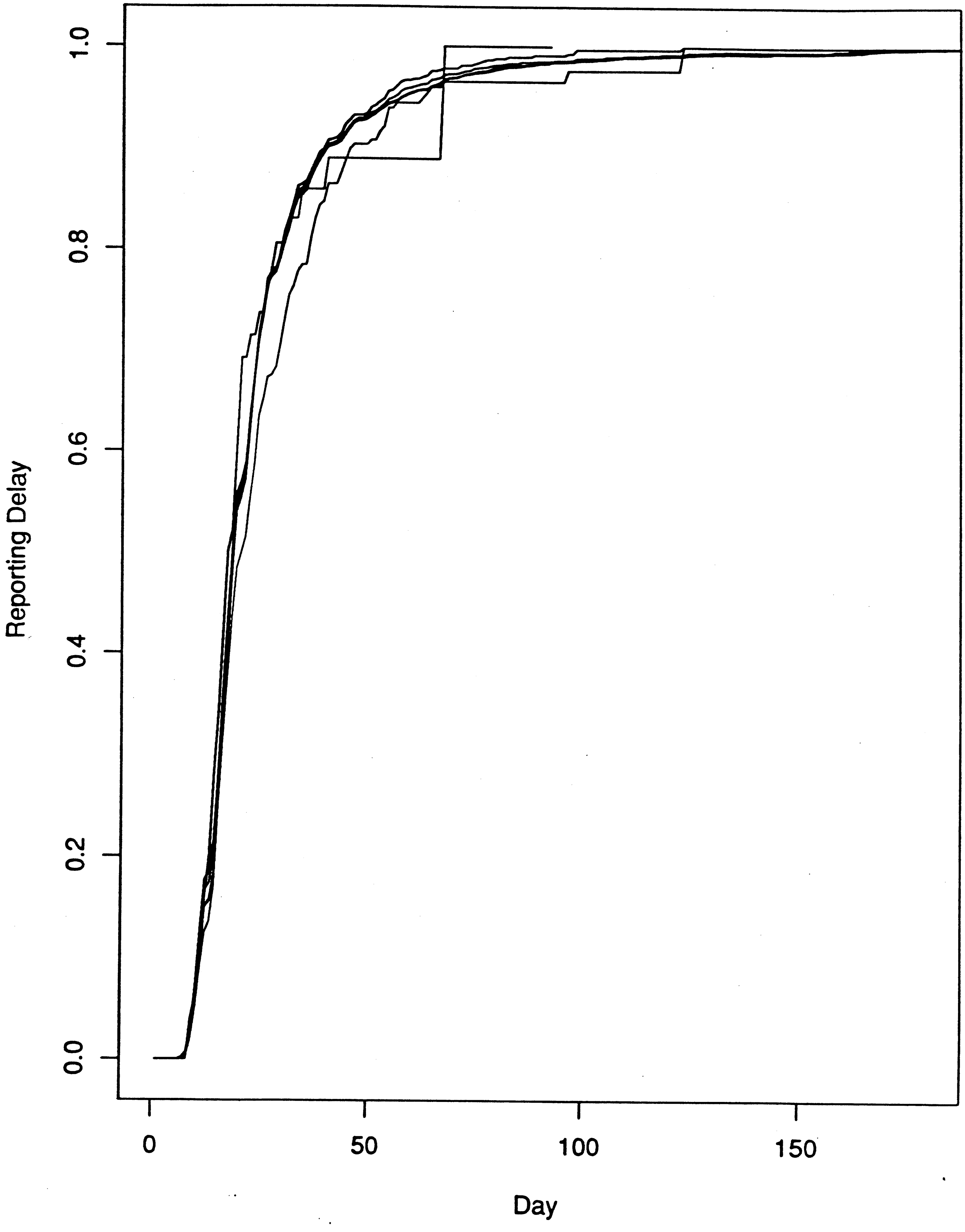
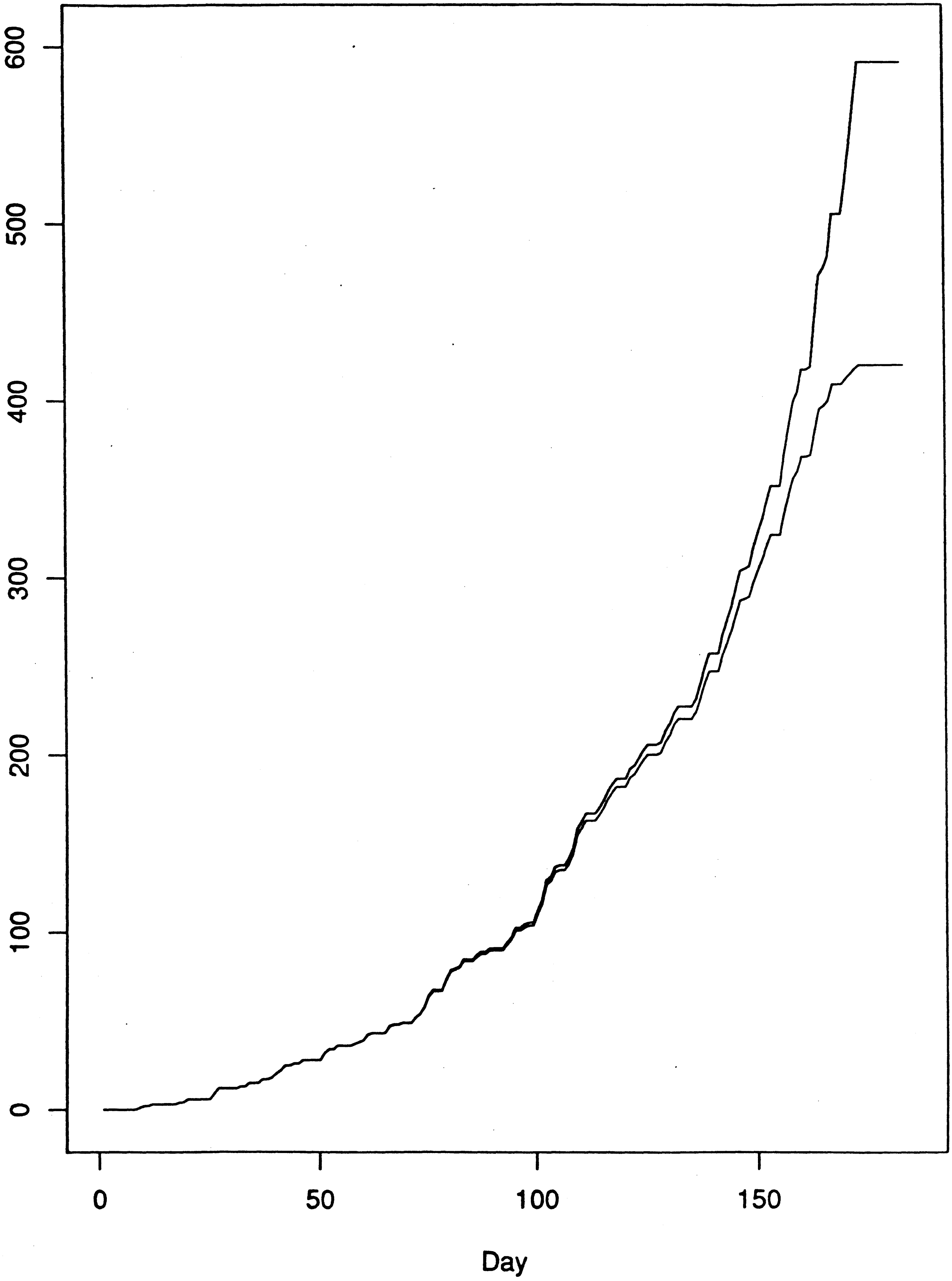


Figure 6. Estimated cumulative number of claims made up to day t (upper curve) and cumulative number reported by day $T = 182$ (lower curve).



TRUNCATED DATA ARISING IN WARRANTY AND FIELD PERFORMANCE STUDIES, AND SOME USEFUL STATISTICAL METHODS

J.D. Kalbfleisch and J.F. Lawless
University of Waterloo
Waterloo, Ontario, N2L 3G1
Canada

SUMMARY

Truncated data arise when a variable is observable only over some portion of its range. In this note we describe how truncated data arise in studies of the field performance or reliability of manufactured items. Failure to account for truncation can lead to biased inferences. We present some useful nonparametric methods, with examples.

Some Key Words: lifetime data, truncated data, nonparametric estimation, field reliability, warranty data