

A Hierarchical Bayesian Approach for Modeling Infant-Mortality and Wearout Failure Modes

Eric Mittman 1*

Department of Statistics, Iowa State University
and

Colin Lewis-Beck

Department of Statistics, Iowa State University

March 28, 2017

Abstract

The text of your abstract. 100 or fewer words.

Keywords: 3 to 6 keywords, (don't reuse words appearing in title)

*The authors gratefully acknowledge Bill Meeker for his comments and suggestions

1 Introduction

1.1 Background

Hard drive failure is the primary cause of data loss and or damage [4]. In addition to the cost of repairing a broken hard drive, lost or corrupted data can have major economic consequences for businesses, researchers, and consumers [6]. To protect against hard drive failure multiple companies now offer backup storage. One of the major online backup companies is Backblaze. While selling storage space is Backblaze’s main business, since 2013 it has been collecting data on hard drives operating at its facility. The purpose is to provide consumers and businesses with reliability information on different hard drive brands and models. The hard drives continuously spin in controlled storage pods where they run until failure. When a hard drive fails it is permanently removed, and new hard drives are regularly added to the storage pods. In addition, the number of storage pods is increasing as Backblaze adds new hard drive brands to the sample. Every quarter Backblaze makes its data publicly available through their website [1]. In addition, Backblaze publishes summary statistics of the different hard drive models currently operating. No other analysis or modeling of the failure data is provided; however, Backblaze does encourage the public to further analyze its data, which for this paper goes through the first quarter of 2016.

1.2 Motivation

The goal of this paper is to model, and compare, the lifetime distribution of the various hard drive brands using parametric lifetime models. There are numerous features of the data, however, that make standard estimation approaches problematic. The first issue is many hard drives have been running for quite some time before entering the sample. Even hard drives starting after 2013 appear in the data after hundreds of hours of operation. Also, although we have four years of data, hard drives rarely fail so many units are still in service. Thus, we have both left truncation and right censoring. The other challenge is each hard drive model varies in number units in operation and total time on test. Some models have hundreds of units in service; others have as few as two. Combined with the truncation this makes stable maximum likelihood (ML) estimates difficult to obtain for many drive models.

Another issue is many product populations contain a mixture of defective and nondefective units. The hazard function for this type of population is often described as a bathtub curve: the beginning of the curve corresponds to defective units failing early, followed by a constant hazard, and then an upswing as units fail from wearout. Ignoring this hazard structure, which is known to exist for computer disk drives, could lead to spurious inference when comparing the reliability of hard drive brand models [2]. A model that accounts for multiple failure modes is the general limited failure population model (GFLP) that combines different failure time distributions; for example, one failure mode for defective units (infant mortality) and a second mode for wearout [2]. Unfortunately, as often the case with lifetime data, the true causes of failure for the hard drives are unknown, which presents challenges for parameter identification in the GFLP model, as discussed in Chan and Meeker (1999). Therefore, we propose to fit the GFLP model using a hierarchical Bayesian approach. With over 60 brand drive models in testing a hierarchical framework is advantageous as it pools information from across brands to get more precise estimates of the individual parameters. This is especially helpful for the Backblaze data as there are many drive models with as few as one or two failures.

1.3 Overview

The structure of the paper is as follows. Section 1 presents the Backblaze data and the unique modeling challenges it presents, such as multiple failure modes, right censoring, and left truncation. Section 2 introduces the GFLP model as a mixture of two Weibull distributions. Using the Backblaze data, in Section 3 we fit a 5 parameter GFLP model that allows for separate Weibull terms for each failure mode and estimates p , the proportion of units subject to infant mortality. In Section 4 we discuss computational and modeling issues that suggests a final model with a common Weibull distribution for defective units and individual Weibull distributions for the wearout failure mode. In Section 5 we compare hard drive brand models based on our final model. In Section 6 we examine the posterior distributions of the model parameters to assess goodness of fit for the hierarchical model. Finally, in Section 7 we review the GFLP model and propose future extensions.

2 Data

As of the first quarter of 2016, Backblaze was collecting data on 63 different hard drive models. Some drive models have been running since 2013, while others have been added at a later date. There are a total of 75,297 hard drives running. The distribution of drives by model varies: some models have only 1 drive in testing; the maximum number of drives running for a specific drive model is 35,860. We can also look at the distribution of total number of failures and total time on test across all drive models. As seen below, some drive models contain a lot of information, while others have few failures or a little time in operation. [INSERT FIGURE 1].

Probability plotting is an effective method to check the adequacy of various parametric models to the data. Identifying whether failure data follows a specific distribution is difficult to do by eye: especially when looking at pdfs or other non-linear plots. However, if we can linearize the cumulative distribution function it is easier to visually assess a distributional goodness of fit; if a model is appropriate, a non parametric estimate of $\hat{F}(t)$ graphed on linearized probability scales should approximately follow a straight line.

For the hard drive data we first estimate the empirical cdf for each model using the Kaplan-Meier estimator [3]. With left truncation, however, the Kaplan-Meier estimate, $\widehat{F}(t)_{KM}$, is conditional based on hard drive i surviving up to time τ_i^L where L is the amount of time the hard drive was running before Backblaze starting monitoring the drive. To get the unconditional distribution we correct the non-parametric estimates using a parametric adjustment outlined by Turnbull, and given in more detail by Meeker and Escobar [7, 5]. For each hard drive model we select τ_{\min}^L , the smallest left truncated time in the sample. Then, using the parametric estimates of the parameters, we calculate $Pr(T > \tau_{\min}^L)$ the probability a hard drive has survived up to τ_{\min}^L . We then apply this correction to the non parametric estimates, which results in the unconditional distribution of time to failure.

In Figure 2 we plot the Kaplan-Meier adjusted cdf for hard drive model X on Weibull paper. Each point on the plot corresponds to a hard drive failure. Censored drives are

not plotted. As mentioned in the introduction, the population of hard drives exhibits two primary failure modes. One mode is a result of manufacturing defects, which cause early failures, known as infant mortality. The second mode is non-defective hard drives that eventually fail due to wearout. Evidence of at least two failure modes is seen in the Kaplan Meier plot with a kink occurring around hour Z. Therefore, fitting a single Weibull model would not be flexible enough to model the failure distribution. [INSERT FIGURE, Do we Want to Put Greenwood Bands on Too?]

3 GFLP Model

3.1 Motivation

We suggest the motivation for the GFLP model in the Data Section. Perhaps expand on this a little more?

3.2 Model

Write out the full Bayes hierarchical model for the GFLP model

3.3 Computation

Discuss the MCMC approach, RStan, etc.

4 Data analysis

4.1 Results

Show some of the fits of the GFLP Model. Also present the unpooled output for a few models. Show parameter estimates with uncertainty. Mention some convergence statistics for the MCMC.

4.2 collapsing μ_1, σ_1

Based on posterior plots, we might fit a simpler model...

4.3 Brand Comparisons

How can we use this model to compare brands? Present comparisons of Quantiles, other parameters. Highlight how this model could be used in an applied sense. Quantiles by brand, etc.

5 Concluding Remarks and Extensions

Review the advantages of fitting the GFLP model and offer future ideas.

SUPPLEMENTARY MATERIAL

Put R Stan code here

References

- [1] Backblaze hard drive data sets. <https://www.backblaze.com/b2/hard-drive-test-data.html>. Accessed: 2016-4-1.
- [2] Victor Chan and William Q Meeker. A failure-time model for infant-mortality and wearout failure modes. *IEEE Transactions on Reliability*, 48(4):377–387, 1999.
- [3] Paul Meier E. L. Kaplan. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.
- [4] Robin Harris. How data gets lost. *ZD Net Storage Bits*, 8 2007.
- [5] W.Q. Meeker and L.A. Escobar. *Statistical Methods for Reliability Data*. Wiley Series in Probability and Statistics. Wiley, 1998.
- [6] David M. Smith. The cost of lost data. *Graziadio Business Review*, 6(3), 2003.

- [7] Bruce W. Turnbull. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3):290–295, 1976.