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 25, 2017

Abstract

The text of your abstract. 100 or fewer words.

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

1 Introduction

1.1 Background

Hard drive failure is the primary cause of data loss and or damage [3]. 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 [4]. 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 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 and 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 generalizes fitting a single failure mode distribution is the general limited failure population model (GFLP) that combines multiple 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 fewer than 5 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 results in us selecting 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 final model and propose extensions of the GFLP model.

2 Data

Introduce the data and show some summary statistics, but NOT box plots. Perhaps show an adjusted Kaplan Meier plots to show a Weibull lifetime distribution could be reasonable.

Possible plots (by drive model): avg. time on test vs. proportion failed total time on test (histogram) total records (by drive model)

3 GFLP Model

3.1 Likelihood

Write out the likelihood for the single Weibull with censoring and right truncation. This can come from the previous paper. Also mention why MLE model alone by brand has its limitations.

3.2 Hierarchical Model

Write out the model, priors, etc for the single Weibull Distribution model.

3.3 Comparison of Approaches

Compare MLE estimates to Weibull and show improved precision due to pooling of information. This can follow the last paper in terms of the plots. I think this section should be shorter, however, since we are just warming up for the GFLP Model.

4 GFLP Model

4.1 Motivation

Point out limitation of single failure mode model (perhaps show a drive model with a kink in the distribution) and then present the GFLP Model. Could also discuss the bathtub hazard. Mention Wayne Nelson.

4.2 Model

Write out the full Bayes hierarchical model for the GFLP model

4.3 Computation

5 Data analysis

5.1 Results

Show some of the fits of the GFLP Model. Show parameter estimates with uncertainty. Mention some convergence statistics for the MCMC.

5.2 collapsing μ_1, σ_1

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

5.3 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.

6 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] Robin Harris. How data gets lost. *ZD Net Storage Bits*, 8 2007.

[4] David M. Smith. The cost of lost data. *Graziadio Business Review*, 6(3), 2003.