2018

A Hierarchical Model for Heterogenous Reliability Field Data

Eric Thomas Mittman
Iowa State University, emittman@iastate.edu

Colin Lewis-Beck
Iowa State University, clewisbe@iastate.edu

William Q. Meeker
Iowa State University, wqmeeker@iastate.edu

A Hierarchical Model for Heterogenous Reliability Field Data

Abstract
When analyzing field data on consumer products, model-based approaches to inference require a model with sufficient flexibility to account for multiple kinds of failure. The causes of failure, while not interesting to the consumer per se, can lead to various observed lifetime distributions. Because of this, standard lifetime models, such as Weibull or lognormal may be inadequate. Usually cause-of-failure information will not be available to the consumer and thus traditional competing risk analyses cannot be performed. Furthermore, when the information carried by lifetime data are limited by sample size, censoring and truncation, estimates can be unstable and suffer from imprecision. These limitations are typical; for example, lifetime data for high-reliability products will naturally tend to be right-censored.

In this paper we present a method for joint estimation of multiple lifetime distributions based on the Generalized Limited Failure Population (GLFP) model. This 5-parameter model for lifetime data accommodates lifetime distributions with multiple failure modes: early failures due to "infant mortality" and failures due to wearout. We fit the GLFP model using a hierarchical modeling approach. Borrowing strength across populations, our method enables estimation with uncertainty of lifetime distributions even in cases where the number of model parameters is larger than the number of observed failures. Moreover, using our Bayesian method, comparison of different product brands is straightforward because estimation of arbitrary functionals are easily computable using draws from the joint posterior distribution of the model parameters. Potential applications include assessment and comparison of reliability to inform purchasing decisions.

Keywords
GLFP, Bathtub hazard, Censored data, Hierarchical models, Bayesian estimation

Disciplines
Statistics and Probability

Comments
This is a manuscript submitted to *Technometrics* (2018).

A Hierarchical Model for Heterogenous Reliability Field Data

Eric Mittman, Colin Lewis-Beck, and William Q. Meeker
Department of Statistics
Iowa State University
Ames, Iowa

January 24, 2018

Abstract

When analyzing field data on consumer products, model-based approaches to inference require a model with sufficient flexibility to account for multiple kinds of failures. The causes of failure, while not interesting to the consumer per se, can lead to various observed lifetime distributions. Because of this, standard lifetime models, such as Weibull or lognormal may be inadequate. Usually cause-of-failure information will not be available to the consumer and thus traditional competing risk analyses cannot be performed. Furthermore, when the information carried by lifetime data are limited by sample size, censoring and truncation, estimates can be unstable and suffer from imprecision. These limitations are typical; for example, lifetime data for high-reliability products will naturally tend to be right-censored.

In this paper we present a method for joint estimation of multiple lifetime distributions based on the Generalized Limited Failure Population (GLFP) model. This 5-parameter model for lifetime data accommodates lifetime distributions with multiple failure modes: early failures (sometimes referred to in the literature as "infant mortality") and failures due to wearout. We fit the GLFP model to a heterogenous population of devices using a hierarchical modeling approach. Borrowing strength across sub-populations, our method enables estimation with uncertainty of lifetime distributions even in cases where the number of model parameters is larger than the number of observed failures. Moreover, using this Bayesian method, comparison of different product brands across the heterogenous population is straightforward because estimation of arbitrary functionals is easy using draws from the joint posterior distribution of the model parameters. Potential applications include assessment and comparison of reliability to inform purchasing decisions.

Keywords: GLFP, Bathtub hazard, Censored data, Hierarchical models, Bayesian estimation

1 Introduction

1.1 Motivation

Consumers have an interest in accurate assessment of product reliability. Toward this end, there usually is a need for models that can accommodate common failure patterns and methods which can make the most of available data and properly account for uncertainty. Doing these things well enables good decision-making in matters related to product life, including purchase decisions and contingency planning.

The reliability of many engineered products follows a similar pattern. Relatively high rates of early failure are due to manufacturing defects. After this "burn-in" period, failure rates stabilize after the majority of defective units have failed. Finally, after prolonged use, rates of failure increase due to wearout. Ignoring this pattern in failure rates can lead to spurious inferences about a product's reliability and suboptimal decisions. This highlights the importance of choosing a sufficiently flexible model.

This paper presents a general framework for statistically modeling reliability field data from a heterogeneous population of components or subsystems operating within a larger population or fleet of systems. Such a model would be useful for applications such as

- Making purchase decisions from among different suppliers of the components or subsystems.
- Making predictions for the number of needed replacement components or subsystems for spare part provisioning.

1.2 Model considerations

Nonparametric methods require relatively few assumptions, but may not be suitable for prediction when extrapolation is required. Moreover, nonparametric estimates made using small samples or data which are heavily censored may be too noisy to be useful. Appropriate parametric models can make better use of available data. For inference, likelihood-based methods are able to deal with truncated and/or censored data. Furthermore, by assuming a parsimonious representation of the data-generating process, parametric models admit

simpler, more intuitive ways to compare sub-populations of interest, borrow strength across sub-populations, and incorporate prior information.

Often no information is available to identify why a unit failed, even when the failed units are available. Physical failure mode analysis (sometimes referred to as "autopsy") can be extremely time consuming and expensive. This presents a dilemma for analysts: the data may indicate multiple failure modes, contraindicating the use of a unimodal parametric distribution. However, without any knowledge of cause of failure, traditional competing risk models are not identifiable. Chan and Meeker (1999) proposed the Generalized Limited Failure Population (GLFP) model. This model can provide a solution to the problem of unknown cause of failure in the special case where there is evidence of two modes of failure which impact different stages of product life. It avoids non-identifiability by introducing a parameter representing the population proportion defective. When this parameter is zero, the GLFP model reduces to a unimodal distribution.

The GLFP model has a meaningful parameterization and accommodates lifetime data with both early and wearout failures. Unfortunately, it can require a lot of data to fit due to the model's complexity. When multiple sources of information are available, partial pooling, accomplished via hierarchical modeling, can reduce the amount of data required to produce stable parameter estimates. When multiple sub-populations are of interest, comparisons based on separate, unrestricted GLFP model fits will be limited to those products with sufficient data. As we will show, hierarchical modeling of the GLFP parameters allows for borrowing of information across sub-populations which imposes "soft" constraints on model parameters. This enables estimation of the lifetime distribution for all sub-populations via shrinkage toward a "pooled" model, with the degree of shrinkage inversely related to the amount of sub-population specific information. We demonstrate the computation of various quantities of interest while comparing reliability across populations using numerical integration over the posterior distribution with samples obtained via Markov Chain Monte Carlo (MCMC).

1.3 Overview

The structure of the paper is as follows. Section 2 introduces the proposed method, potential applications, and examines previous related work. Section 3 introduces the GLFP model and illustrates its use with a subset of hard drive failure data from the company, Backblaze. In Section 4, we describe the complete Backblaze Hard Drive data set, which is used throughout as a motivating example. In Section 5 we discuss using hierarchical modeling of the GLFP model parameters to extend the GLFP model to multiple sub-populations. Section 6 applies the hierarchical GLFP model to the hard drive data. First, we do model selection among four models of increasing complexity, using approximate leave-one-out cross-validation. Next, for the selected model, we assess the model fit using a graphical comparison of replicated data sets generated from the posterior distribution to the observed data. Finally, we present an evaluative comparison of sub-populations taking into account practical considerations related to product lifetime. Section 9 describes potential applications of our model and limitations for inference implied by the model assumptions.

2 Background

In engineering applications, a product can often fail from one out of a set of possible malfunctioning components. For example, a computer system can break if the mother board, disc drive or power supply stop working. Circuit boards (CB) can fail due to a manufacturing defect or later as a result of wearout of certain components. The general name for such products is a series system where the lifetime of the product is the minimum failure time across different components or risks (Nelson, 1982, Chapter 5). A common assumption in series system modeling is the time to failure for the different risks are statistically independent. Thus, the overall reliability of a unit is the product of the reliability of each of the risks. Then parameter estimation is straightforward if the cause of failure is known for each observation. However, in many situations the cause of failure is unknown to the analyst. This is referred to in the literature as masking.

Previous papers have employed various assumptions and methods to model masked lifetime failure data. When modeling computer system failures Reiser et al. (1995) assumed

each observed failure came from a known subset of failure modes, and estimation was performed using a Bayesian approach. Chan and Meeker (1999) labeled the cause of circuit board failures as early (due to defects), unknown, or wearout based on the time of observed failures. This helped identify parameters when using maximum likelihood (ML) estimation. Extending Chan and Meeker's analysis, Basu et al. (2003) performed a Bayesian analysis with informative priors to better identify early versus late failure modes without making any assumptions about the causes of failure. Berger and Sun (1993) introduced the Poly-Weibull distribution where the failure time is the minimum of a several Weibull distributions, corresponding to different failure modes. Ranjan et al. (2015) considered a competing risk model for early and wearout failures as a mixture of Weibull and exponential distributions. Treating the unknown failure modes as incomplete data, an expectation maximization (EM) algorithm providing ML estimates was used, in addition to Bayesian estimation.

3 Model for field data

3.1 Weibull distribution parameterization

The Weibull cumulative distribution function (cdf) is

$$\Pr(T \leq t|\alpha,\beta) = F(t|\alpha,\beta) = 1 - \exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right], \quad t > 0, \tag{1}$$

where $\beta > 0$ is the Weibull shape parameter and $\alpha > 0$ is a scale parameter. Because $\log(T)$ has a smallest extreme value distribution (a member of the location-scale family of distributions), the Weibull cdf can also be written as

$$\Pr(T \leq t|\mu,\sigma) = F(t|\mu,\sigma) = \Phi_{SEV}\left[\frac{\log(t)-\mu}{\sigma}\right], \quad t > 0,$$

where $\Phi_{SEV}(z) = 1 - \exp[-\exp(z)]$ is the standard smallest extreme value distribution cdf and $\mu = \log(\alpha)$ and $\sigma = 1/\beta$ are, respectively, location and scale parameters for the distribution of $\log(T)$. Therefore, the Weibull distribution is a member of the log-location-scale family of distributions.

We will use an alternative parameterization where α is replaced by the p quantile $t_p = \alpha\,[-\log(1-p)]^{\sigma}$. Replacing α in (1) with the $\alpha = t_p/[-\log(1-p)]^{\sigma}$ gives

$$\begin{aligned}\Pr(T \le t|t_p,\sigma) = F(t|t_p,\sigma) &= 1 - \exp\left[-\left(\frac{t}{t_p/[-\log(1-p)]^{\sigma}}\right)^{1/\sigma}\right] \\ &= 1 - \exp\left[\log(1-p)\left(\frac{t}{t_p}\right)^{1/\sigma}\right], \quad t > 0.\end{aligned}$$

There are two important reasons for using this parameterization.

- Especially with a high-reliability product, it will be easier to elicit prior information about a quantile (t_p) in the lower tail of the distribution than it will be to elicit prior information about α (approximately the 0.63 quantile). In addition, there is generally available information about the shape parameter, σ, for a given failure mechanism (e.g., if the failure is due to a wearout mechanism, then it is known that $\sigma < 1$).

- Because of heavy censoring in reliability field data, the parameter estimates of the μ and σ parameters will generally be highly correlated and thus specification of independent marginal prior distributions would be be inappropriate. On the other hand, estimates of t_p and σ, for some appropriately chosen value of p, will be approximately independent, allowing the easier elicitation and specification of independent marginal prior distributions. For example, if a typical field data set has 10% of units failing, then choosing $t_{0.05}$ would work well.

- Bayesian MCMC estimation will be better behaved due to reduced correlation between t_p and σ (relative to α and σ).

3.2 Generalized Limited Failure Population model

Let F_1, F_2 be Weibull distributions with parameters (t_{p1}, σ_1) and (t_{p2}, σ_2), respectively. The Generalized Limited Failure Population model (GLFP) of Chan and Meeker (1999) is defined as follows: Let $T \sim \text{GLFP}(\pi, t_{p1}, \sigma_1, t_{p2}, \sigma_2)$. Then

$$\Pr(T \le t) = H(t) = 1 - (1 - \pi\, F_1(t))(1 - F_2(t)),\ t > 0,\ 0 < \pi < 1.$$

The GLFP model can be understood as a mixture model with a binary latent variable, $\delta_i \overset{ind.}{\sim} \text{Bernoulli}(\pi)$. δ_i is an indicator for whether unit i is defective or not (i.e., susceptible to an early failure). Expressed conditional on δ_i,

$$\begin{aligned} P(T \leq t|\delta_i = 1) =& 1 - (1 - F_1(t))(1 - F_2(t)) \\ P(T \leq t|\delta_i = 0) =& F_2(t). \end{aligned}$$

The parameter π represents the proportion of units susceptible to early failure, and hence susceptible to both failure modes. Here the cause of failure is not assumed to be known, thus units from the same population are exchangeable.

Taking the derivative of the (marginal) cdf, the density for the GLFP model is

$$\begin{aligned} h(t|\pi, t_{p1}, \sigma_1, t_{p2}, \sigma_2) = \pi f_1(t|t_{p1}, \sigma_1) \left(1 - F_2(t|t_{p2}, \sigma_2)\right) + \\ f_2(t|t_{p2}, \sigma_2) \left(1 - \pi F_1(t|t_{p1}, \sigma_1)\right). \end{aligned}$$

Note that the p quantile of the GLFP model may be found by setting $p = H(t_p)$ and solving numerically for t_p.

3.3 Censoring and truncation

A common feature of lifetime data is right-censoring. In the analysis of reliability field data, it is rare that all units are observed until failure. If a unit has not yet failed when the data are analyzed it is considered right-censored. In other words, right-censoring puts a lower bound on the failure time.

When an observation is left-truncated, it would not have been observed if it had failed prior to a particular time, which we refer to as the left-truncation time. Left-truncation is a common feature of observational lifetime data, where the factors leading to inclusion in the data set are uncontrolled and/or the population of interest has a history prior to any data collection. Ignoring left truncation can lead to biased estimates. However, dropping left truncated observations should be avoided because it could substantially reduce the total available information. Therefore, we incorporate both right censoring and left truncation into the likelihood.

3.4 Likelihood

We now give the general form for the likelihood function, taking into account left truncation and right censoring. Let t_i denote the end of the observed lifetime of unit i, in hours. Let t_i^L be the left truncation time, the age of unit i when data reporting commenced. Additionally, let c_i be an indicator for censoring; $c_i = 1$ if the failure time is right-censored, $c_1 = 0$ if the unit failed (at time t_i). The likelihood for the GLFP is a function of the parameters $\boldsymbol{\theta} = (\pi, t_{p1}, \sigma_1, t_{p2}, \sigma_2)$. Assuming the lifetimes of all units are independent, the likelihood for the data, $t_1, \ldots, t_n$ is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\frac{h(t_i; \boldsymbol{\theta})}{1 - H(t_i^L; \boldsymbol{\theta})} \right]^{1-c_i} \left[\frac{1 - H(t_i; \boldsymbol{\theta})}{1 - H(t_i^L; \boldsymbol{\theta})} \right]^{c_i}$$

4 Motivating example

4.1 Backblaze hard drive data

Backblaze is a company that offers cloud backup storage to protect against data loss. 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-models. The hard drives continuously spin in controlled storage pods. Drives are run until failure. When a hard drive fails it is removed and replaced. In addition, the number of storage pods is increasing as Backblaze adds drives to its storage capacity. Every quarter Backblaze makes its hard drive data publicly available through their website (`https://www.backblaze.com/b2/hard-drive-test-data.html`, accessed January 18, 2018). In addition, Backblaze publishes summary statistics of the different drive-models currently operating. No other analysis or modeling of the failure data is provided, other than simple rates and proportions. Backblaze does, however, encourage others to further analyze their data.

As of the first quarter of 2016, Backblaze was collecting and reporting data on 63 different drive-models. Some drive-models have been running since 2013 or before, while others were added at a later date. Data have been reported on 75,297 different hard drives that are or were in operation. The number of drives varies by drive-model; some drive-

models only have a service record for a single drive whereas the maximum number of daily service records for a single drive-model is 35,860. Figure 1 shows a scatterplot of the total observed running time in hundred of thousands hours versus the total number of failures for drive-models with at least 3 failures. For model identification, a minimum of three failures was the criterion for inclusion of drive-models into our analysis.

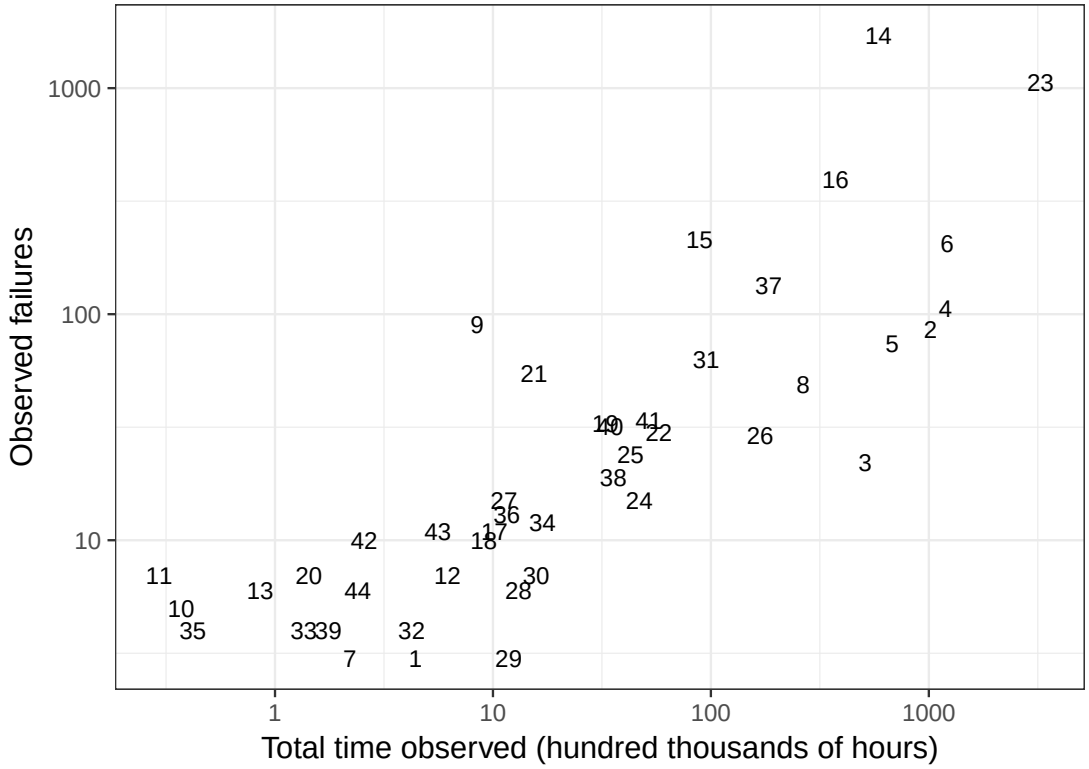


Figure 1: Scatterplot of total observed running time (hundred thousands of hours) versus total number of observed failures. Each number labels a unique drive-model, numbered between 1 and 44. Axes are on the log scale.

4.2 Analysis of a single drive-model

To illustrate the GLFP model, we will present an analysis of the drive-model with the most observed failures, Drive-model 14. The Kaplan-Meier (K-M) estimate plotted below suggests at least two failure modes. The curve has a slow ramp up in failures until about 10,000 hours (early failures), and increases more rapidly from about 10,000 to 20,000 hours (presumably wearout failures). Besides our empirical observation, it is well known that

computer hard drives have a mixture of early and late failures (Chan and Meeker, 1999). Therefore, there is both an empirical and a theoretical justification to apply the GLFP model.

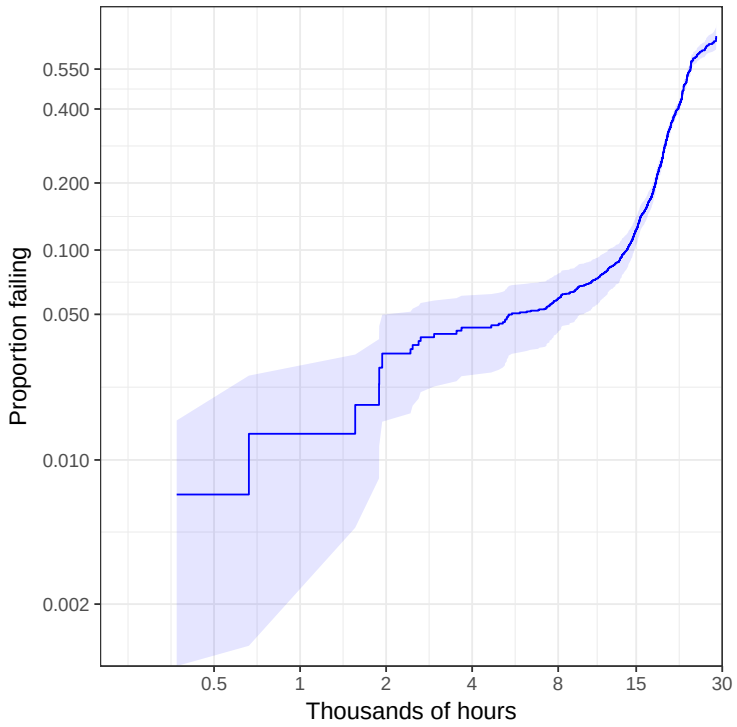


Figure 2: K-M estimate for Drive-model 14. Uncertainty band correspond to pointwise standard errors (using Greenwood's formula.)

To estimate the parameters of the GLFP model we use a Bayesian approach, selecting proper, but generally diffuse, prior distributions to improve identification of the model parameters. Having a proper prior distribution also ensures a proper posterior distribution. We reparameterize the component Weibull distributions using the 0.50 quantile for the early failure mode ($t_{0.5,1}$) and the 0.20 quantile for the wearout failure mode ($t_{0.2,2}$).

When eliciting prior distribution information it is much easier to ask about recognizable characteristics of a distribution instead of the parameters of the distribution. Following the approach used in (Meeker et al., 2017, Section 15.2.2) we will use diagonal braces ($\langle ., . \rangle$) to refer to 95 percent central probability intervals, rather than the standard model parameters when specifying prior distributions. Using this convention the prior distributions used in

our analyses are

$$t_{0.5,1} \sim \text{Lognormal}\langle 1.7,\, 7.6 \times 10^6 \rangle, \quad t_{0.2,2} \sim \text{Lognormal}\langle 8.6,\, 5.6 \times 10^7 \rangle$$

$$\pi \sim \text{Logit-normal}\langle 1.0 \times 10^{-3},\, 7.1 \times 10^{-1} \rangle, \quad \sigma_1, \sigma_2 \overset{ind.}{\sim} \text{Lognormal}\langle 7.4 \times 10^{-3},\, 1.3 \times 10^2 \rangle.$$

These prior distributions put probability mass on a wide range of values for all model parameters — much larger than we would expect for a typical Weibull distribution. Thus, we consider these prior distributions to be relatively uninformative.

Table 1 gives the posterior median and 95% credible intervals for the 5 GLFP parameters for Drive-model 14. The parameter π is an estimate of the proportion of drives susceptible to early failure. As expected, this proportion is small with a median value of 0.05. The shape parameter estimates for the two Weibull distributions also match intuition. The early failure mode puts posterior probability on a values of β_1 less than 1, which corresponds to a decreasing hazard. Conversely, the credible interval for β_2, the wearout mode, is strictly above 1, implying an increasing hazard function. The two quantiles are also well identified with the early failure mode having an much earlier time to reach the 0.50 quantile compared to the time to reach the 0.20 quantile for the wearout mode.

	2.5%	50%	97.5%
π	0.033	0.054	0.099
$\beta_1 = 1/\sigma_1$	0.47	1.13	1.72
$\beta_2 = 1/\sigma_2$	4.47	4.70	4.95
$t_{0.5,1}$	1.03	2.28	3.99
$t_{0.2,2}$	18.0	18.2	18.6

Table 1: Posterior 95% Credible Intervals for the 5 GLFP parameters for Drive-model 14. Quantiles are in thousands of hours.

Probability plotting is a simple method to assess and compare the adequacy of members of the log-location-scale family of distributions. After properly transforming the axes of a plot, graphing an empirical estimate of fraction failing as a function of time along with pointwise confidence bands provides a visual check for distributional goodness of fit. We applied this method using the Kaplan-Meier nonparametric estimate of the empirical cdf.

With left truncation, however, the standard Kaplan-Meier estimator is biased, so we used an adjusted version due to Meeker and Escobar (1998, Chapter 11) (see Appendix for a detailed description).

In Figure 3, we overlay the posterior median of the fitted GLFP model onto the adjusted K-M estimate with axes on the Weibull probability scale. The plot also contains 90% pointwise credible bands associated with each estimate. While the Weibull model is inadequate for these data, the GLFP model fits quite well, as it is able to adequately describe the rapid increase in the empirical cdf between 8000 and 20,000 hours.

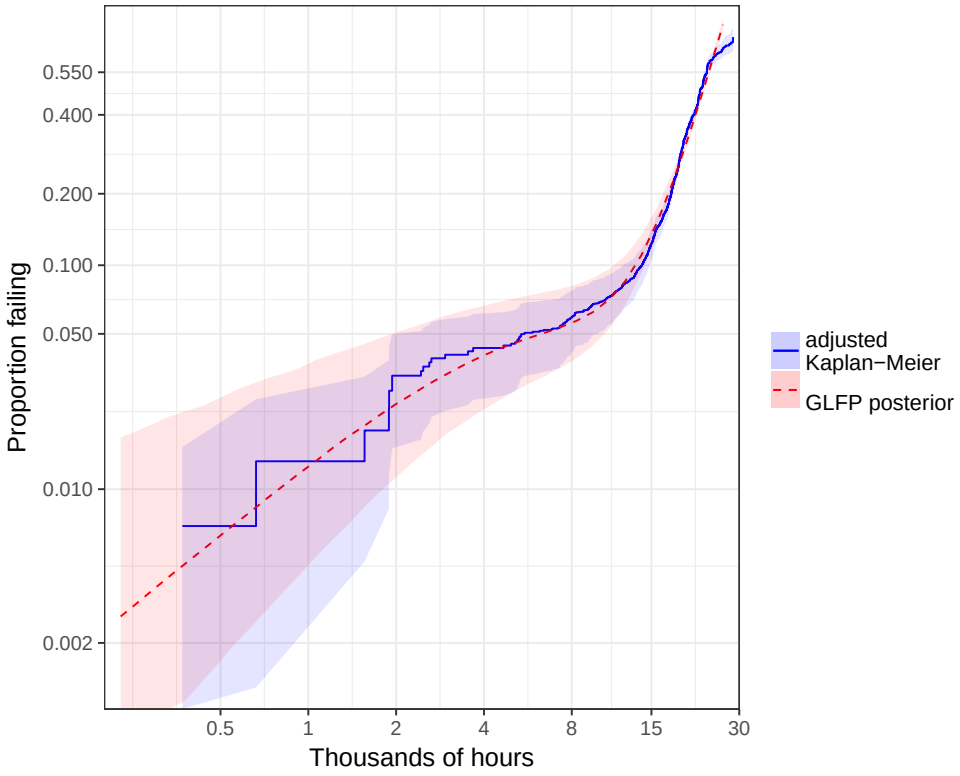


Figure 3: The estimated GLFP for Drive-model 14 plotted on Weibull paper. The dashed line corresponds to the median of the posterior draws; pointwise 90% credible bounds are shown in red. The solid line is an adjusted K-M estimate with pointwise 90% credible bounds in blue.

5 Hierarchical GLFP model

5.1 Conditional likelihood

In order to describe an entire product population, consisting of different but similar sub-populations, and because of the limited amount of data from many of the sub-populations, we model sub-population-specific parameters hierarchically, borrowing strength across sub-populations. Let

$$T_{ig} \overset{ind.}{\sim} \text{GLFP}\left(\pi_g, t_{p_1 g}, \sigma_{1g}, t_{p_2 g}, \sigma_{2g}\right), \tag{2}$$

where $g = 1, \ldots, G$ indexes the sub-populations. The likelihood for the hierarchical GLFP model is a function of the sets of parameters $\boldsymbol{\theta_g} = (\pi_g, t_{p_1 g}, \sigma_{1g}, t_{p_2 g}, \sigma_{2g})$, one set for each sub-population, g. Assuming the lifetimes of all units are independent within and across sub-populations, and conditional on the fixed parameters, the likelihood for the data, $i = 1, \ldots, n_g$, is given by

$$L(\boldsymbol{\Theta}) = \prod_{g=1}^{G} \prod_{i=1}^{n_g} \left[\frac{h(t_{ig}; \boldsymbol{\theta_g})}{1 - H(t_{ig}^L; \boldsymbol{\theta_g})}\right]^{1-c_{ig}} \left[\frac{1 - H(t_{ig}; \boldsymbol{\theta_g})}{1 - H(t_{ig}^L; \boldsymbol{\theta_g})}\right]^{c_{ig}}$$

where T_{ig} is the observed failure or survival time of unit i in sub-population g; t_{ig}^L is the left truncation time for unit i in sub-population g; and c_{ig} is an indicator if unit i in sub-population g is right censored. All times are again in hours.

5.2 Parameter distributions

In a hierarchical model, the parameters are modeled as random variables, varying across different sub-populations. For example

$$\begin{gathered}
\sigma_{1g} \overset{ind.}{\sim} \text{Lognormal}\left(\eta_{\sigma_1}, \tau_{\sigma_1}^2\right) \\
\sigma_{2g} \overset{ind.}{\sim} \text{Lognormal}\left(\eta_{\sigma_2}, \tau_{\sigma_2}^2\right) \text{Tr}\left(0, 1\right) \\
t_{p_1 g} = \exp\left(\mu_{1g} + \sigma_{1g}\, \Phi^{-1}(p_1)\right) \overset{ind.}{\sim} \text{Lognormal}\left(\eta_{t_{p_1}}, \tau_{t_{p_1}}^2\right) \\
t_{p_2 g} = \exp\left(\mu_{2g} + \sigma_{2g}\, \Phi^{-1}(p_2)\right) \overset{ind.}{\sim} \text{Lognormal}\left(\eta_{t_{p_2}}, \tau_{t_{p_2}}^2\right) \\
\pi_g \overset{ind.}{\sim} \text{Logit-normal}(\eta_\pi, \tau_\pi^2).
\end{gathered} \tag{3}$$

As discussed in Section 3.1, we reparameterize the Weibull distribution in terms of a quantile and the shape parameter because observed failure data are often more informative about the lower tail of their respective lifetime distribution (note that if T has a Weibull distribution, σ is a shape parameter for the distribution of T and a scale parameter for the distribution of $\log(T)$). We truncate the distribution of σ_{2g} at 1 (indicated by $\text{Tr}(0,1)$ above) restricting the wearout failure mode to have an increasing hazard function (i.e., $\beta_{2g} = 1/\sigma_{2g} > 1$).

5.3 Priors

The complete GLFP model requires prior distributions on the hierarchical parameters. We suggest weakly informative proper distributions that allow the data to overwhelm the prior and provide information about the distribution of the population parameters. Prior knowledge about product lifetimes is useful for choosing the mean of the location parameter (η) distributions. However, the corresponding standard deviations should be diffuse to allow the location to vary depending on the observed data. For the scale parameters (τ) we recommend half-Cauchy distributions as suggested by Gelman et al. (2014).

5.4 Nested models

In principle, we prefer the full model (3) because we tend to believe that every population has a distribution with a distinct set of parameters. In practice, however, we may consider restrictions to reduce the number of parameters if the data do not support their inclusion. We considered four models, all based on (3), differing by which model parameters are held constant across sub-populations. These are (from most to least restrictive):

Model 1 $\pi_g = \pi, \quad t_{p_1 g} = t_{p_1}, \quad \sigma_{1g} = \sigma_1, \quad t_{p_2 g} = t_{p_2}, \quad \sigma_{2g} = \sigma_2$

Model 2 $\pi_g = \pi, \quad t_{p_1 g} = t_{p_1}, \quad \sigma_{1g} = \sigma_1, \quad \sigma_{2g} = \sigma_2$

Model 3 $\pi_g = \pi, \quad t_{p_1 g} = t_{p_1}, \quad \sigma_{1g} = \sigma_1$

Model 4 $t_{p_1 g} = t_{p_1}, \quad \sigma_{1g} = \sigma_1$

This set of model specifications is chosen based on the data, interpretation of the model, as well as estimation considerations. Sub-population-model-specific parameters for the the wearout failure mode $(t_{p_2 g}, \sigma_{2g})$, and the proportion defective (π_g), are considered as a means to account for heterogeneity across sub-populations in the right tails of the failure distribution. Going from a common model for all sub-populations and gradually increasing the complexity of the model, Model 4 is the most flexible. Model 4 allows for the probability of early failure as well as the shape and scale parameters for the wearout failure mode to vary across sub-populations.

For all of the nested models, the parameters for the early failure mode are common across sub-populations. We found that there was often insufficient information to model these parameters hierarchically. Moreover, assuming a common distribution for early failures provides a meaningful interpretation and comparison of π_g and $\pi_{g'}$ $(g \neq g')$. Also the failure-time distribution of the defective subpopulation (as opposed to the proportion defective) is not of high interest.

5.5 Model selection

From the point of view of a consumer or purchaser, the goal of an analysis such as this is may be to rate manufacturers and/or to inform future purchases based on predicted product performance. In this case, we should choose the model that will provide the most accurate prediction for future units of previously observed products. One appropriate criterion for model selection, then, is the log pointwise predictive density (lpd),

$$\begin{aligned} \text{lpd} &= \sum_{g=1}^{G} \sum_{i=1}^{n_g} \log[p(t_{new,g,i} | t, t^L, c)] \\ &= \sum_{g=1}^{G} \sum_{i=1}^{n_g} \log \left[\int p(t_{new,g,i} | \theta) p(\theta | t, t^L, c) d\theta \right]. \end{aligned}$$

Using this criterion, we should choose the model with the highest expected lpd (elpd) for a new data set $\{t_{new,g,i} : g = 1, \ldots, G; i = 1, \ldots, n_g\}$. Notationally suppressing the

conditioning on t^L and c,

$$\text{elpd} = \text{E}_h \, \text{lpd} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \int \log[p(t_{new,g,i}|t)] h(t_{new,g,i}) dt_{new,g,i}, \tag{4}$$

where h is a density for the true distribution of $t_{new,g,i}$. Although h is unknown, leave-one-out cross-validation provides a robust estimator:

$$\widehat{\text{elpd}} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \log[p(t_{g,i}|t_{-(g,i)})]. \tag{5}$$

The `R` package, `loo` (Vehtari et al., 2016), computes an approximation of (5) provided that the observations t_i are conditionally independent. It employs an importance sampling method that uses S saved MCMC draws to compute smoothed importance weights, $w_{i,s}$, approximating $\log[p(t_{g,i}|t_{-(g,i)})]$ with $\sum_{i=1}^{S} w_{i,s} p(t_{g,i}) / \sum_{s=1}^{S} w_{i,s}$. In addition to providing an estimate of (4), their method also produces a standard error based on an asymptotic result. For details, refer to Vehtari et al. (2017).

6 Model for the Backblaze hard drive data

We now apply this general methodology to the hard drive failure data from Backblaze. There are 44 different drive-models in the data set, and significant heterogeneity in the number of observed failures, time observed (Figure 1), and the parameter estimates of the underlying GLFP. The hierarchical GLFP allows us to model the entire drive-model population, borrowing information across drive-models, allowing estimation for those drive-models for which little information exists.

6.1 Prior distributions

Following the recommendations in Section 5.3, we specify the following prior distributions for each of the four nested GLFP models. Different sets of restrictions require different prior distribution specifications, which are assigned as follows:

Model 1 Constrain all drive-models to the same GLFP distribution. For this "reduced" model we assume the same prior distributions as in Section 4.2.

Model 2 $t_{p_2 g}$ varies by drive-model. To help with model identifiably, we tighten the priors on the early failure mode:

$$\pi \sim \text{Logit-normal}\langle 7.0 \times 10^{-3},\, 2.6 \times 10^{-1} \rangle,$$

$$\sigma_1 \sim \text{Lognormal}\langle 1.4 \times 10^{-1},\, 7.1 \rangle,$$

$$t_{p_1} \sim \text{Lognormal}\langle 2.2 \times 10^{1},\, 5.5 \times 10^{4} \rangle.$$

Model 3 $t_{p_2 g}$ and σ_{2g} vary by drive-model. Prior distributions for the constrained parameters are the same as for Model 2.

Model 4 π_g, $t_{p_2 g}$ and σ_{2g} vary by drive-model. Prior distributions for constrained parameters are the same as for Model 2.

Where applicable, prior distributions on hyperparameters are as listed below. We use half-Cauchy prior distributions on hierarchical scale parameters. As for the location hyperparameters, we choose weakly informative prior distributions centered around the corresponding prior mean for the non-hierarchical model.

$$\begin{aligned}
\eta_{\pi} &\sim \text{Normal}(-3, 1) \\
\tau_{\pi} &\sim \text{Half-Cauchy}(0, 1) \\
\eta_{\sigma,2} &\sim \text{Normal}(0, 2) \\
\tau_{\sigma,2} &\sim \text{Half-Cauchy}(0, 1) \\
\eta_{t_{.2_2},2} &\sim \text{Normal}(9, 2) \\
\tau_{t_{.2_2},2} &\sim \text{Half-Cauchy}(0, 1).
\end{aligned}$$

6.2 Model fitting

Each model was fit using the `rstan` (Stan Development Team, 2016) package in `R` (R Core Team, 2013), which implements a variant of Hamiltonian Monte Carlo (HMC) (Betancourt and Girolami, 2015). HMC jointly updates all model parameters by simulating energy preserving paths with random initial momentums along the posterior density. This is done

to reduce autocorrelation and efficiently explore the posterior. Multiple chains were run, each with 1500 iterations after 1500 warmup iterations: 4 chains were run for Models 1, 2 and 3 for a total of 6,000 post burn-in iterations and 16 chains were run for Model 4 for a total of 24,000 post burn-in iterations. The Gelman-Rubin potential scale reduction factor was used to provide a check for adequate mixing of the multiple chains. Upon convergence, $\hat{R}$, converges to 1. The respective maximum values of $\hat{R}$ across all model parameters for Models 1 through 4 (5, 50, 95 and 140 parameters, respectively) were 1.005, 1.028, 1.009 and 1.004. Other diagnostics provided by the software (tree-depth, and divergent transitions) did not indicate problems with sampling. Plots of the posterior draws were inspected for parameters with the fewest effective samples; these did not suggest features that were inadequately explored.

6.3 Model selection

As we have discussed, while we prefer to allow all of the GLFP model parameters to be drive-model specific, there are practical issues with fitting the full model. For example, due to heavy left-truncation, there may be insufficient information for particular drive-models to estimate all of the parameters.

Figure 4 provides a visual comparison for Models 1 through 4 by plotting the pointwise posterior median of the cdf for each drive-model. The large left panel shows the adjusted K-M estimate. The smaller plots correspond to Models 1 through 4. To make the plots comparable, we use an adjustment based on the parametric model (Model 4). Note that while all the paths for the posterior estimates necessarily pass through the origin, the K-M estimates do not, because, for the K-M estimator, inference is entirely conditional on survival up to left-most left-trunctation time.

These plots suggest that Model 1 is insufficient to explain the observed data. Model 2 certainly captures more of the heterogeneity observed in the K-M estimates; we observe that the assumption of a common σ_2 results in very similar progressions in the cdf, which may be too restrictive. Model 3, which allows σ_2 to vary by sub-population, suggests there is evidence to support variation in this parameter among drive-models. While there are differences between Model 3 and 4, they are relatively subtle.

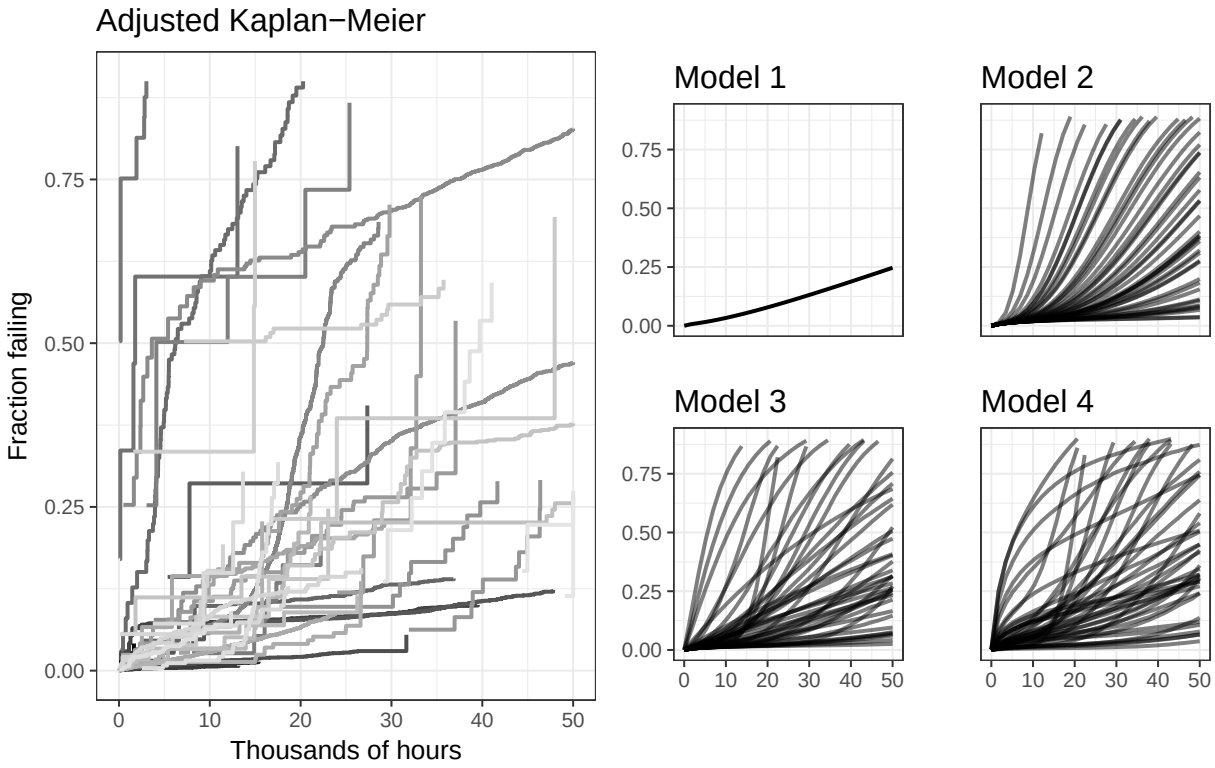


Figure 4: Left: Adjusted K-M estimates of the time to failure for each of the drive-models in the Backblaze data. Right: Pointwise posterior median cdf estimates for Models 1-4, ordered left to right, top to bottom.

We can also compare posterior estimates of time to failure under Models 1-4 for each drive-model individually. For these comparisons we calculate pointwise the posterior median of the proportion failing over a grid. That is, for every model, and for a fixed set, $\tilde{t} = \{\tilde{t}_1, \ldots, \tilde{t}_M\}$, we compute

$$\text{median}\left\{\text{H}\left(\tilde{t}_m | \pi_g^{(s)}, t_{p_1 g}^{(s)}, \sigma_{1g}^{(s)}, t_{p_2 g}^{(s)}, \sigma_{2g}^{(s)}\right); s = 1, \ldots, S\right\}, \tag{6}$$

where $\theta^{(s)}$ is posterior sample s from $p(\theta|y)$.

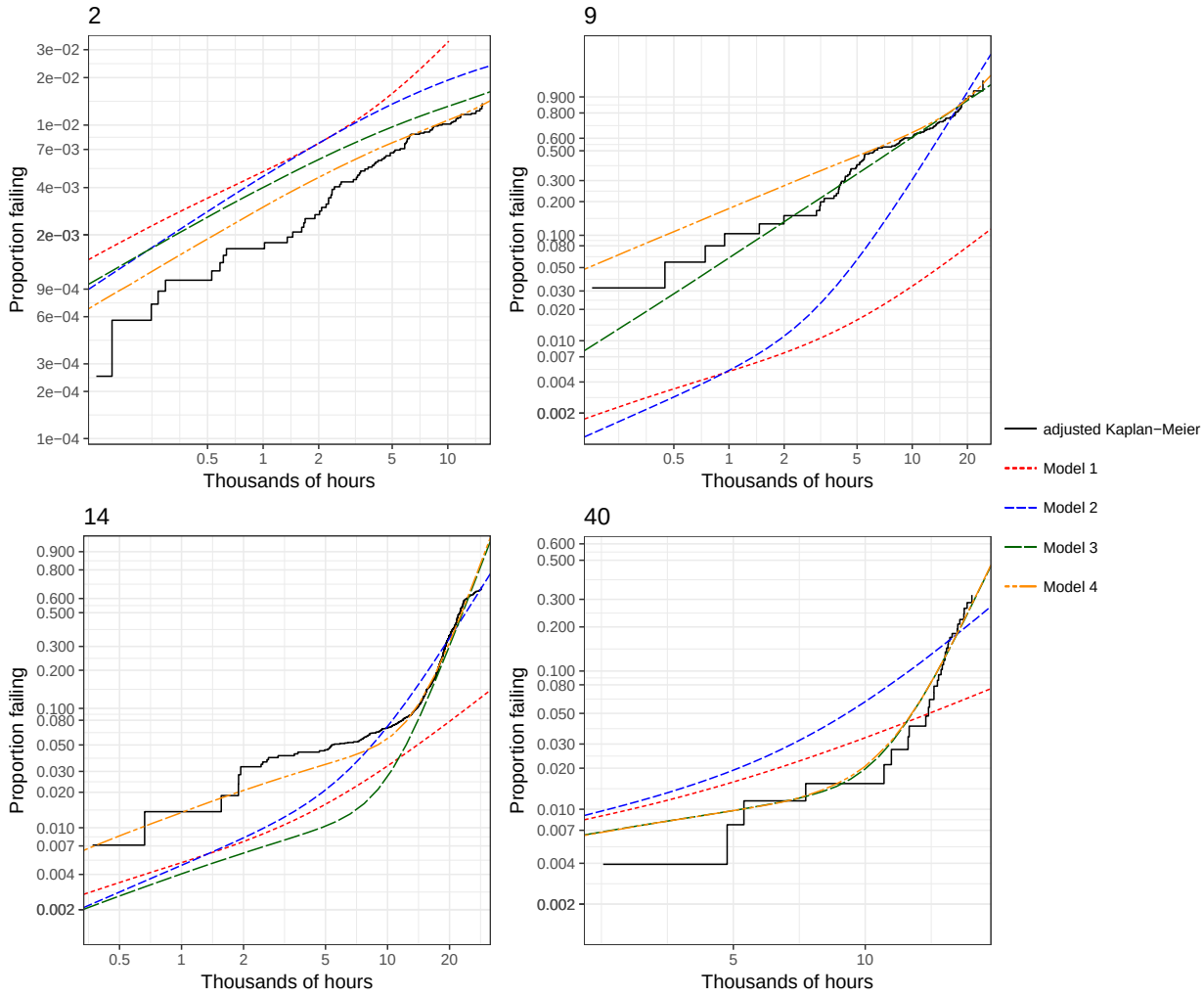


Figure 5: Posterior median of the proportion failing as a function of time under Models 1-4 for sample of drive-models, with axes scales as for Weibull probability plotting. The solid step function is the adjusted K-M estimate.

The GLFP fit for a representative selection of drive-models (Drive-models 2, 9, 14, and 40) are presented in Figure 5 (See the Supplementary Material for similar plots for the other drive-models). The black step functions again correspond to the adjusted K-M estimates. As model complexity increases, the hierarchical-model GLFP curves become more flexible and are better able to describe the observed failure data. All of the GLFP models appear to produce higher estimates of the proportion failing, relative to K-M, for Drive-model 2. This behavior is not unexpected, however, because hierarchical models improve model stability through shrinkage; in this context, shrinkage is toward an average cdf. Model 4 is

	$\widehat{\text{elpd}}$	Difference in $\widehat{\text{elpd}}$	SE of the Difference
Model 4	-13309.5	40.7	11.3
Model 3	-13350.2	458.8	31.0
Model 2	-13809.0	3674.6	96.2
Model 1	-17483.6		

Table 2: Expected Log Pointwise Predictive Density (elpd) for each model specification. Each model is compared to the more parsimonious model below. The estimated difference of expected leave-one-out prediction errors between the two models, as well as the standard error of the difference, is also presented.

in closest agreement.

Drive-model 9 shows a high proportion of early failures, a fact which Model 1 fails to capture. Model 2 addresses this by estimating an earlier time to wearout, but still results in a poor fit to the data. Models 3 and 4 conform more closely to the adjusted K-M estimate. The disagreement between Models 3 and 4 for Drive-model 9, is interesting: Model 4 allows the proportion of defective units to vary by drive-model, but Model 3 appears to fit just as well in this case, by increasing the scale parameter for the wearout failure mode.

For Drive-model 14, Models 1-3 produce lower estimates of the proportion of drives failing, and for Drive-model 40, they produce higher estimates. On the other hand, Model 4 follows the K-M closely across stages of lifetime. The differences between Models 3 and 4 and Models 1 and 2 are visually apparent. Differentiating between Model 3 and 4 is less clear; for instance, the estimates from Model 3 and 4 are nearly identical for Drive-model 40.

Results showing the estimated elpd and associated standard errors for all 4 models are shown in Table **??**. We calculated the difference in expected predictive accuracy for Model 4 versus 3, Model 3 versus 2, and Model 2 versus 1, as well as the standard error of the difference (Table 2). For the models we considered, each successive increase in model complexity increases elpd. Of all the models, Model 4 has significantly better elpd compared to the the other 3 models.

7 Model Assessment

7.1 A simulation-based approach

Having selected a model, it is good practice to check that the model provides an adequate representation of the data-generating mechanism. One way to do this is to draw replicate data sets from the posterior predictive distribution

$$p(t_{rep,g}|t) = \int f(t_{rep,g}|\theta_g)\, p(\theta_g|t)\, d\theta_g.$$

This is a general approach recommended by Gelman et al. (1996), which works when no classical goodness-of-fit test is available. When the model is adequate for the data, the predictive distribution of a finite set of data characteristics — chosen to reflect features relevant to the goals of the analysis — should not contradict the corresponding characteristics observed in the actualized data.

There are many possible approaches to the posterior predictive check. We propose the comparison of K-M estimates for each sub-population derived from replicated data drawn from the posterior predictive distribution to the K-M estimate based on the actualized data. This graphical approach does not provide a test statistic, but provides rich feedback with respect to the fitted model's agreement with the data.

Because the variability in the K-M estimator is primarily a function of the number of at-risk units, we want to replicate the pattern of censoring that was observed. That is, for unit i of sub-population g,

$$p(t_{rep,g}) = \int \prod_{i=1}^{n_g} f(t_{new}|t^L_{g,i}, c_{g,i}, \theta)\, p(\theta|t) d\theta.$$

Because we are working with posterior draws, $\theta^{(s)}$, we can draw from this distribution by choosing a draw, $\theta^{(s)}$, at random, then conditionally draw $t_{rep,g}$ from $f(t_{new}|t^L_{g,i}, c_{g,i}, \theta^{(s)})$.

An obstacle remains, owing to the fact that the censoring time is unknown for observed failures. Our admittedly simple approach is to assume that failed units would have been the censored had they survived until the latest censoring or failure time for that drive model. Our method for generating replicate data sets, $t^{(s)}_{rep,g}$, $s = 1, \ldots, R$, for each sub-population $g = 1, \ldots, G$, is as follows:

For $s = 1, \ldots, R$, do:

1. Sample r uniformly from $\{1, \ldots, S\}$.

   For $i = 1, \ldots, n_g$, do:

2. Simulate $\tilde{t}^{(s)}_{rep,g,i} \sim \text{GLFP}(\pi^{(r)}_g, t^{(r)}_{p1}, \sigma^{(r)}_1, t^{(r)}_{p2g}, \sigma^{(r)}_{2g})$.

3. Repeat step 2 until $\tilde{t}_{rep,g,i} > t^L_{g,i}$.

4. Set $d^{(s)}_{g,i} = \begin{cases} \max\{t_{g,i} : i \geq 1\}, & \text{if } c_{g,i} = 0 \\ t_{g,i}, & \text{if } c_{g,i} = 1. \end{cases}$

   Lastly, set $(t^{(s)}_{rep,g,i}, c^{(s)}_{rep,g,i}) = \begin{cases} (d^{(s)}_{g,i}, 1), & \text{if } \tilde{t}^{(s)}_{rep,g,i} > d^{(s)}_{g,i} \\ (\tilde{t}^{(s)}_{rep,g,i}, 0), & \text{if } \tilde{t}^{(s)}_{rep,g,i} \leq d^{(s)}_{g,i}. \end{cases}$

Finally, plot K-M estimates based on the R replicated data sets along with the one based on the actualized data on the same axes. Discrepancies between the replicated and actual data can be used to diagnose inadequacies of the model.

7.2 Application to the Backblaze data

For the hard-drive data, we used the posterior samples obtained in fitting Model 4 to generate $R = 19$ replicate data sets as described above. The resulting plots are displayed in Figure 6.

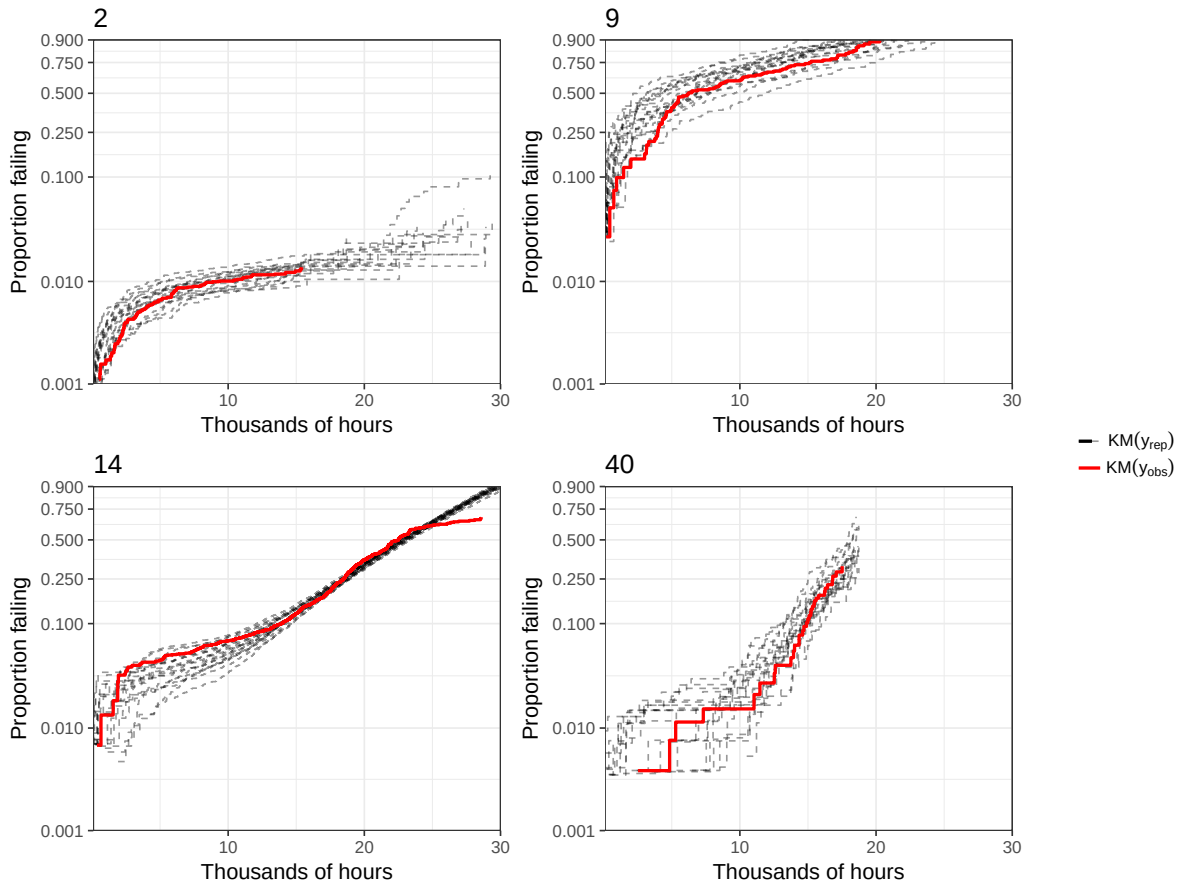


Figure 6: Adjusted K-M estimates of the proportion failing for a representative subset of drive-models. Both the original data (**bold** line) and 19 "replicated" data sets from the posterior predictive distribution (*dashed* lines) are shown.

Only a representative sample of plots is shown here for brevity (see the Supplementary Material for similar plots for all of the drive-models). For most drive-models, the estimates based on the observed data look similar to those based on the replicated data. There are a few drive-models, however where this is not the case. Drive-model 14 shows a discrepancy between the GFLP fit and the K-M fit in the right tail. In particular, the GLFP predictions appear conservative, overpredicting the proportion failing after about 2.5 years. Due to the large number of failures for this drive-model, the posterior for the hierarchical model is similar to the stand-alone model. Other drive-models showing a lack of fit are 4 and 34. We first address Drive-model 4; our diagnosis of 34 is analogous, so the following comments extend to that drive-model as well.

To diagnose the cause of the discrepancies between the observed data for Drive-model 4 and its posterior predictive distribution, we consider the event plot for a random sample of units for this Drive-model while looking at estimates of lifetime (Figure 7). From these plots, we can see that the discrepancy between the posterior and K-M estimates is driven by failures among a relatively small subset of units, namely those with left-truncation times smaller than 2500 hours. Further investigation reveals that, if we exclude the newest 75 drives, those with truncation times less than 800 hours, the K-M estimate falls within the posterior 90% pointwise credible band. This situation suggests that, despite belonging to the same drive-model, there may be something different about the newer drives leading to worse reliability.

A possible explanation for the discrepancy shown in Figure 7 is that the units of Drive-model 4 were introduced into the field over a period of 22 months. Manufactures of hard drives (and most other products) often make changes to the product design over time to either improve reliability or to reduce cost. Such changes will sometimes have an important effect on the product's failure-time distribution, invalidating one of our important assumptions that failure-time distribution for a drive-model does not depend on the date of manufacture.

Figure 8 shows a similar pair of plots for Drive-model 43. The data for this Drive-model are heavily truncated; the observed units had run for quite some time before data were made available with the earliest left-truncation time at 12,189 hours. In the bottom half of Figure 8 we contrast the posterior median for drive-model 43 to an average GLFP cdf for the entire population, which we call the "global average;" a description of the global average is given in Appendix B. We can see that the lack of data for this drive-model during early life results in a diffuse posterior centered around the global average until about 10,000 hours when they diverge. Although the posterior is close to the adjusted K-M, it is shrunk toward the global average.

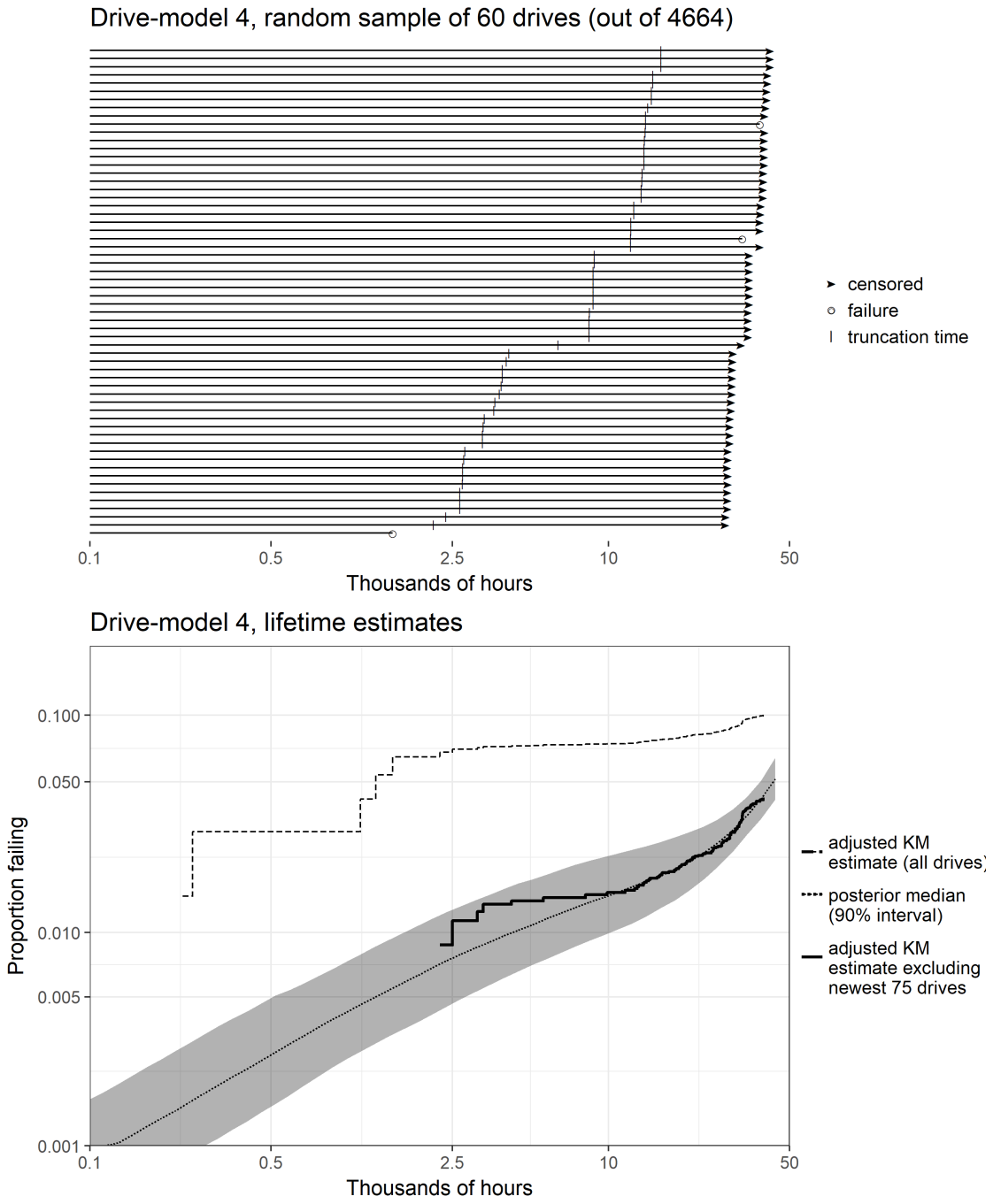


Figure 7: Top: Event plot for 60 randomly sampled drives of Drive-model 4. Note that all drives in the sample are left truncated well past 1000 hours. Bottom: The adjusted K-M estimate (dashed step function) is substantially higher than the pointwise posterior median (dashed line; shaded region showing 90% credible interval). However, this discrepancy is due to several early failures among the small set of drives with the earliest left-truncation times (the newest set). The same adjusted K-M estimator after the exclusion of the newest 75 drives (solid step function) shows close agreement with the posterior median.

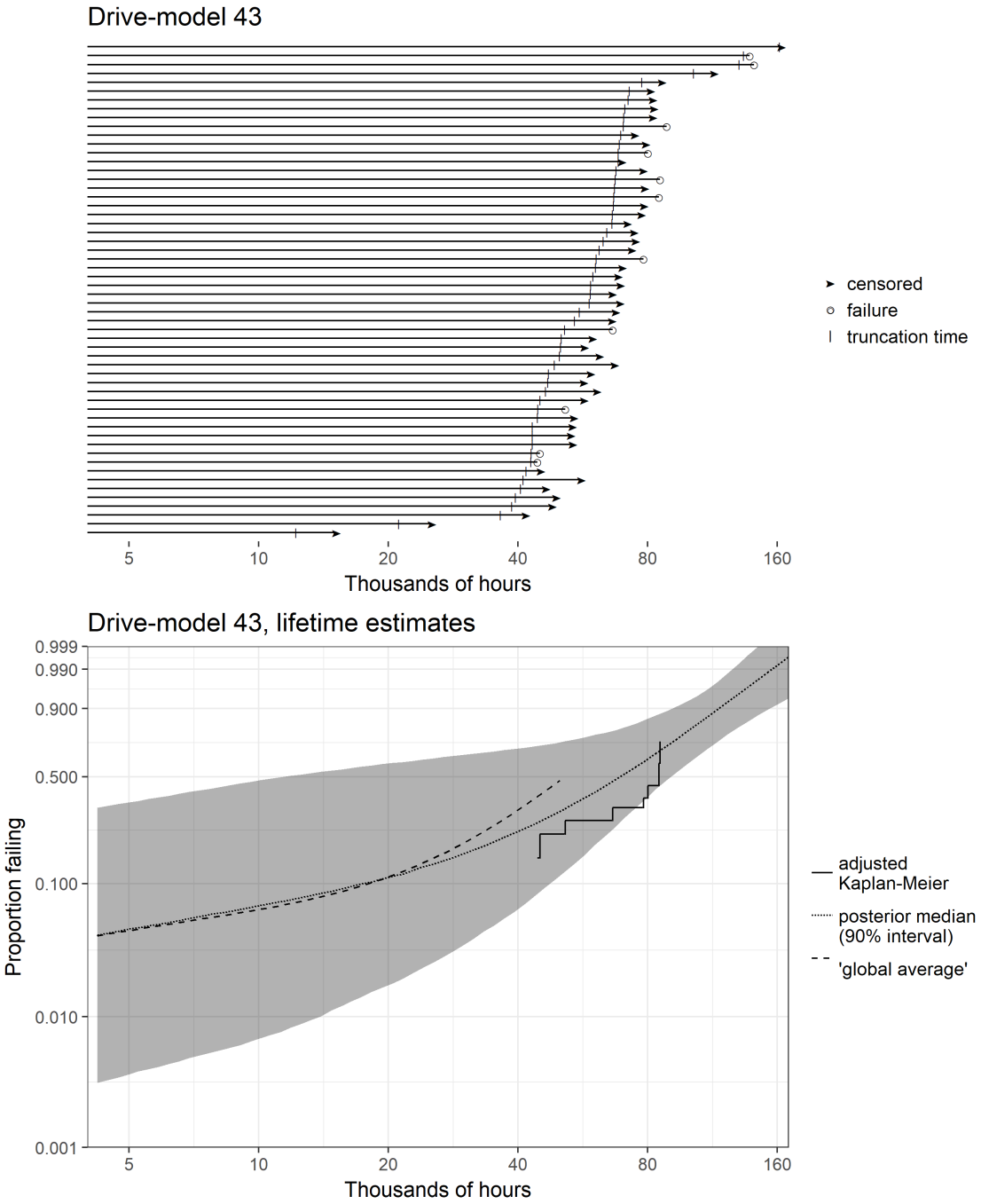


Figure 8: Top: Event plot for all units of Drive-model 43. Data for all units are left-truncated, being observed conditional on survival beyond 12,000 hours. Bottom: The dashed step function correponds to the adjusted K-M estimate (solid line). Due to the heavy left-truncation, the posterior median for Drive-model 43 (dotted line) coincides with the posterior median of the "global average" (dashed line) until the first left-truncation time.

8 Comparing Sub-populations

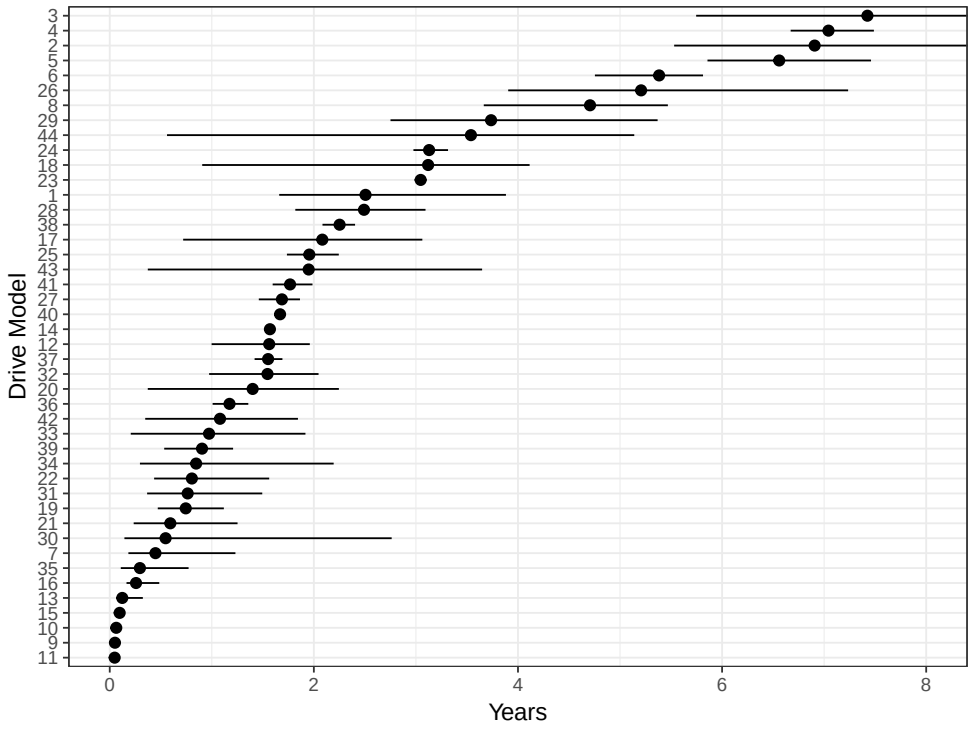
8.1 Comparisons based on model parameters

As discussed in Section 3.1, we use a parameterization of the Weibull distribution in terms of a chosen quantile, making prior elicitation straightforward. Posterior sampling immediately gives marginal posterior distributions for quantiles, which in reliability applications is typically more informative than a measure of central tendency. To compare the unconditional p quantile for each sub-population, plots of posterior credible intervals may be used.

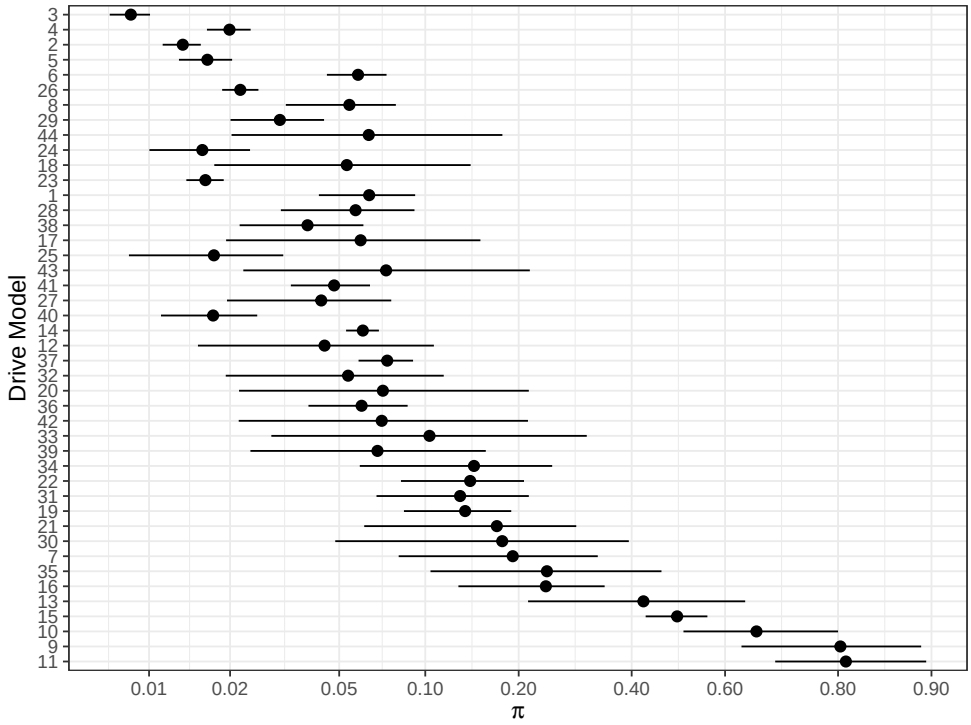
Depending on the set of model restrictions used, simlar plots may be used to compare the p quantiles for non-defective units ($t_{p_2 g}$) or the proportion defective (π_g).

8.2 Example: Drive-model comparison based on quantiles and proportion defective

The top panel of Figure 9 shows estimates of $t_{0.10}$ (i.e., the amount of time it takes for 10% of hard drives to fail, i.e. the quantile corresponding to the $H(t)$ distribution in Section 3.2), for all drive-models. In this application 50% CIs were used due to considerable posterior uncertainty in some of the parameters: especially for those with few drives in operation. Another interesting feature to compare is π_g, the proportion defective. The bottom panel of Figure 9 shows the posterior credible intervals of π_g for each drive-model. The plot is again sorted according to $t_{0.10}$ so it can be compared to the plot above. While for many of the drive-models the ordinal ranking is the same as on the top, there are some drive-model comparisons, for example 23 and 18, that differ in ranking if compared using $t_{0.10g}$ or π_g.



(a) Point estimate and 50% CI for $t_{0.10}$



(b) Point estimate and 50% CI for π

Figure 9: (a) 50% CI (in years) for $t_{0.10}$ (b) 50% CI for π plotted on the logit scale. Both plots are sorted based on the median value of the marginal posterior distribution of $t_{0.10}$.

8.3 Comparisons based on posterior predictive distribution

By utilizing samples from the full posterior distribution, in addition to estimation of the model parameters, we can also produce forecasts for new units or groups of units. In many situations, this may more directly address issues of importance.

The *posterior predictive* distribution incorporates two sources of variability: the posterior uncertainty in the parameters, which largely depends on the amount of data collected, and the uncertainty in future observations conditional on those parameters.

For example,

$$p(t_{g,new}|t) = \int p(t_{g,new}|\theta_g)p(\theta_g|t_g)\, d\theta_g$$

gives the posterior predictive density for a new unit from sub-population g.

We can sample from this distribution by drawing $t_{g,new}^{(s)}$ from $\text{GLFP}(\pi_g^{(s)}, t_{p1}^{(s)}, \sigma_1^{(s)}, t_{p2g}^{(s)}, \sigma_{2g}^{(s)})$, for $s = 1, ..., S$, using the saved posterior draws for the model parameters.

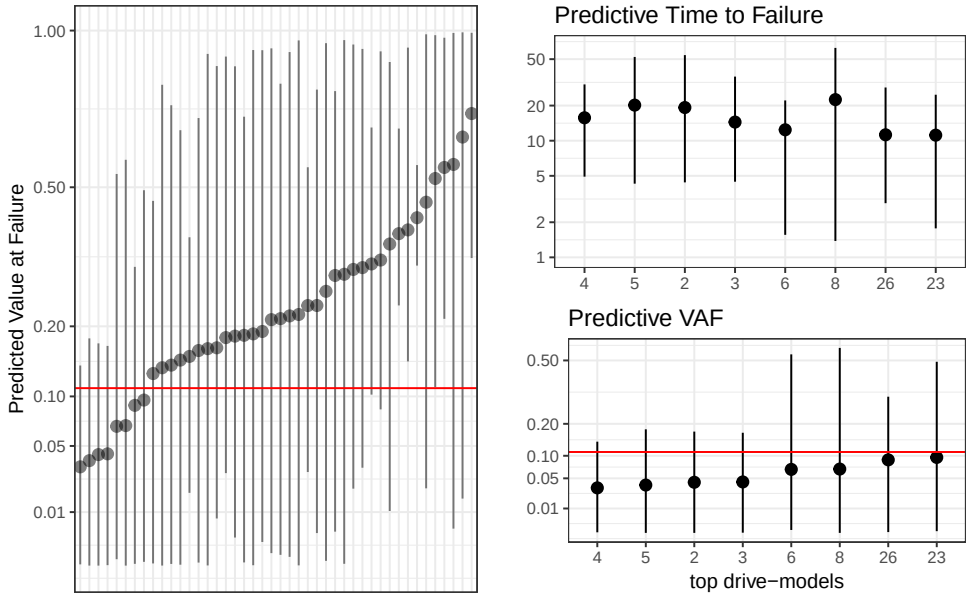


Figure 10: Left: 90% CI of the posterior predictive distributions for $e^{-0.4t_{g,new}}$ the depreciated value at failure (VAF) for a new unit for each Drive-model. Right: A comparison (90% CI) of the predictive distribution of TTF (top) and VAF (bottom) based on 40% for best 11 drive-models, shows that rankings produced by the two measures are not equivalent.

8.4 Example: Drive-model comparison based on predictive distribution of new unit

We consider the problem of ranking drive-models with respect to economical value. From a business perspective, it is clear that we should prefer the drive-models which will provide more years of service. For ease of exposition, we will assume that the purchase price of a hard-drive is the same across drive-models.

The mean time-to-failure (MTTF) for drive-model d can be estimated by

$$\frac{1}{S}\sum_{s=1}^{S} t_{g,new}^{(s)}.$$

While MTTF is often an important metric, due to the anticipation of advancements in technology, we can expect that the relative value of computer hardware will depreciate rather quickly. In simple terms, given two hard drives, one would prefer to have both of them work for five years than for one to work for nine years and the other to work for just one year. To account for depreciation, rather than using MTTF as the metric for drive-model comparison, we use the value at replacement. The US Internal Revenue Service considers computer hardware to be "5-year" equipment (US Department of Treasury, 2016). Using"the declining balance method" the rate of depreciation is 40% per year.

Let $L(t) = e^{-0.4t}$ represent the value at the time of failure relative to a new unit. We can rank the drive-models by

$$\mathrm{E}(L(t_{g,new})|t)) \approx \frac{1}{S}\sum_{s=1}^{S} L(t_{g,new}^{(s)}).$$

Posterior credible intervals for $L(t_{g,new})$ are shown in the left panel of Figure 10. A zoomed version of the top eight drive-models, as ranked by $\mathrm{E}(L(t_{g,new})|y)$ are shown in the bottom right panel. The panel above shows credible intervals for MTTF. We can see that ranking by MTTF would lead to a different result – a new Drive-model 8 unit is expected to last the longest, but the possibility of it failing early is higher than for the drive-models ranked highest by our depreciation-aware metric. The top four drive-models (4, 5, 2, 3) seem comparable, with a new Drive-model 4 unit expected to depreciate the most. However, given the discrepancy of the model fit for Drive-model 4 (noted in the previous

subsection), in the absence of more early life data, we might harbor some lingering concern about possible quality-control issues with this Drive-model.

9 Discussion

This paper offers a new approach to modeling and making inference for the lifetime of consumer products using incomplete field reliability data. Products with long lifetimes often have few failures and when failures do occur the cause is frequently unknown to the analyst. The GLFP model provides a framework to accurately model data with evidence of multiple failure modes. Moreover, when there are multiple product populations within the same product class, our hierarchical modeling approach borrows strength across brands with many observed failures to help make inference for those populations with little information. This can be important when trying to leverage multiple sources of data with various sample sizes and levels of censoring and truncation. We found that the use of weakly informative priors on the hyperparameters can increase the overall efficiency of the MCMC. We now address some important model assumptions and their potential for impacting statistical inference.

First, we assume the failures occurring from various causes can be approximately assigned to two phases of product life. In practice, there might be additional phases, clearly supported by the data, making the GLFP model too rigid. This rigidity is illustrated by the right tail of the estimated distribution in Figure 2.

Second, we assume exchangability of units within a sub-population. This assumption is due in part to ignorance about which units are defective but also of potentially important covariates. We do not account for batch effects or varying conditions in the facility which have impacted the observed failure rates. These factors are confounded with drive-model in our analysis. An example of a possible batch effect is noted in Section 7.1 and illustrated by Figure 7.

Consumers of manufactured goods usually lack the data and expertise to perform a failure analysis to determine the cause of failure. Moreover, their interest is primarily in the lifetime distribution of a product rather than the specific cause of failure. The ability to fit marginal models, such as GLFP, allow consumers and manufacturers to accurately

model products with bathtub hazard lifetime distributions and make comparisons across different product sub-populations. Because the Bayesian approach provides full posteriors distributions once a model is fit, the analyst can easily estimate a functional, such as a quantile or depreciation factor, that allow products to be compared using statistics that are meaningful in the context of the actual application.

While not presented here, another potential application is forecasting the number of future failures over a fixed period of time for a current population of drives. Hong et al. (2009) proposed a method, based on the Poisson-binomial distribution. It would be straightforward to adapt their approach to work with our method. For details and examples see Hong et al. (2009, Sect. 6) and Xu et al. (2015).

Acknowledgment

The authors express their gratitude to Nick Clark for his helpful comments and suggestions.

A Truncation adjusted Kaplan-Meier estimate of lifetime

We first start with a nonparametric estimate of the empirical cdf for each sub-population using the Kaplan-Meier estimator. With left truncation, however, the standard Kaplan-Meier estimator for drive-model g, denoted by $\widehat{F_g(t)}_{KM}$, is conditional on survival up to $t^L_{g,\min}$, the shortest reported running time of all units in of sub-population g for which records are available. To produce unconditional estimates, we adapt the adjustment method outlined by Meeker and Escobar (1998, Chapter 11). For each sub-population we select $t^L_{g,\min}$, the smallest left truncated time in the sample. By sampling from the full posterior distribution, because $\Pr(T > t^L_{g,\min}|\theta_g)$ (the probability that a hard drive has survived up to $t^L_{g,\min}$) is a function of the model parameters, we can easily compute its posterior median, $\widehat{A}_{\text{med}} = \widehat{\Pr}(T > t^L_{g,\min}|\theta_g)$. We compute the adjusted estimate by

$$\widehat{F(t)}_{KMadj} = \widehat{A}_{\text{med}} + \left(1 - \widehat{A}_{\text{med}}\right) \widehat{F_g(t)}_{KM},\ t > t^L_{g,\min}.$$

While this adjustment is may be negligible for sub-populations with little truncation, in our Backblaze example, five drive-models receive upward adjustments of greater than 5 percent and the estimated time to failure distribution of one drive-model (30) was adjusted by nearly 16 percent, in part because the shortest truncation time for all observed units was approx. 2.3 years.

B Definition of global average for Model 4

In Section 7.1, to illustrate the concept of shrinkage in our hierarchical model, we refer to a "global average" which represents an average GLFP cdf for the entire population, $\mathrm{H}\left(\cdot|\bar{\pi}, \mu_1, \sigma_1, \bar{\mu}_2, \bar{\sigma}_2\right)$. Since μ_1 and σ_1 are common across all sub-populations in our model, these can already be interpreted as "global." For the parameters that vary across sub-populations, we select values corresponding to the medians of the hierarchical distributions (3) conditional on the hyperparameters. In particular, we set

$$\bar{\pi} = \text{logit}^{-1}(\eta_\pi),\ \bar{\mu}_2 = \eta_{t_2} - m_{\sigma_2}\Phi^{-1}(.2) \text{ and } \bar{\sigma}_2 = m_{\sigma_2}.$$

Let $J(\cdot|a,b)$, $J^{-1}(\cdot|a,b)$ denote the cdf and inverse cdf, respectively, for a ognormal distribution with log-location parameter a and log-scale parameter b. Then

$$m_{\sigma_2} = J^{-1}[0.5 \cdot J(1|\eta_{\sigma_2}, \tau_{\sigma_2})|\eta_{\sigma_2}, \tau_{\sigma_2}],$$

which is the median of a ognormal distribution with parameters η_{σ_2}, and τ_{σ_2}, truncated to the interval $(0, 1)$.

We estimate the global average pointwise, using draws from the joint posterior distribution, $H\left(\tilde{t}|\eta_\pi^{(s)}, \mu_1^{(s)}, \sigma_1^{(s)}, \eta_{t_2}^{(s)}, \eta_{\sigma_2}^{(s)}, \tau_{\sigma_2}^{(s)}\right)$, $s = 1, 2, \ldots, S$. The computation is thus similar to that shown in (6).

C Supplementary Material

Model code and plots Stan code for Model 4 as well as the complete set of plots for Figures 5 and 6 (supplementary.pdf)

Backblaze data set Lifetime data set for all 44 Backblaze drive-models used in our analysis. (backblaze_hd_data.csv)

References

Backblaze (Accessed January 18, 2018), "Backblaze hard drive data sets." `https://www.backblaze.com/b2/hard-drive-test-data.html`.

Basu, S., Sen, A., and Banerjee, M. (2003), "Bayesian Analysis of Competing Risks with Partially Masked Cause of Failure," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 52, 77–93.

Berger, J. O. and Sun, D. (1993), "Bayesian analysis for the poly-Weibull distribution," *Journal of the American Statistical Association*, 88, 1412–1418.

Betancourt, M. and Girolami, M. (2015), "Hamiltonian Monte Carlo for hierarchical models," *Current Trends in Bayesian Methodology with Applications*, 79–101.

Chan, V. and Meeker, W. Q. (1999), "A failure-time model for infant-mortality and wearout failure modes," *IEEE Transactions on Reliability*, 48, 377–387.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014), *Bayesian Data Analysis*, vol. 2, Chapman & Hall/CRC Boca Raton, FL.

Gelman, A., Meng, X.-L., and Stern, H. (1996), "Posterior predictive assessment of model fitness via realized discrepancies," *Statistica Sinica*, 6, 733–760.

Hong, Y., Meeker, W. Q., and McCalley, J. D. (2009), "Prediction of Remaining Life of Power Transformers Based on Left Truncated and Right Censored Lifetime Data," *The Annals of Applied Statistics*, 3, 857–879.

Meeker, W. and Escobar, L. (1998), *Statistical Methods for Reliability Data*, Wiley Series in Probability and Statistics, John Wiley & Sons Hoboken, NJ.

Meeker, W., Hahn, G., and Escobar, L. (2017), *Statistical Intervals: A Guide for Practitioners and Researchers, Second Edition*, Wiley Series in Probability and Statistics, John Wiley & Sons, Hoboken, NJ.

Nelson, W. B. (1982), *Applied Life Data Analysis*, John Wiley & Sons, Hoboken, NJ.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Ranjan, R., Singh, S., and Upadhyay, S. K. (2015), "A Bayes analysis of a competing risk model based on gamma and exponential failures," *Reliability Engineering & System Safety*, 144, 35–44.

Reiser, B., Guttman, I., Lin, D. K., Guess, F. M., and Usher, J. S. (1995), "Bayesian inference for masked system lifetime data," *Applied Statistics*, 44, 79–90.

Stan Development Team (2016), "RStan: the R interface to Stan," `http://mc-stan.org`, R package version 2.14.1.

US Department of Treasury, I. (2016), "Instructions for Form 4562," `https://www.irs.gov/pub/irs-pdf/i4562.pdf`.

Vehtari, A., Gelman, A., and Gabry, J. (2016), "loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models," `https://CRAN.R-project.org/package=loo`, R package version 1.1.0.

— (2017), "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC," *Statistics and Computing*, 27, 1413–1432.

Xu, Z., Hong, Y., and Meeker, W. Q. (2015), "Assessing risk of a serious failure mode based on limited field data," *IEEE Transactions on Reliability*, 64, 51–62.