VALOR MORGHULIS: BAYESIAN PREDICTION OF CHARACTER DEATHS IN A SONG OF ICE AND FIRE

Noah Gale, Siyu Zhang and Hanbo Sun
Department of Statistics
University of Michigan, Ann Arbor
{noahgale, zhsiyu, hanbosun}@umich.edu

1 Introduction

'Valor Morghulis' translates from the constructed language of Valerian to 'All Men Must Die,' and in George R.R. Martin's acclaimed A Song of Ice and Fire (ASoIaF) series, life is short and brutish enough to make it true. Almost one third of all characters introduced in ASoIaF have died by the end of the most recent book, with the author showing no signs of slowing down.

Several other Bayesian papers have been written on ASoIaF, including a Bayesian survival analysis paper [9], and a Bayesian prediction Analysis of the Point of View character chapters for the forthcoming sixth book in the series [8].

We will attempt to predict the chance characters will live or die using Bayesian Methods on the Weibull, Generalized Linear, and Hierarchical Generalized Linear Models. We believe that the number of chapters a character has survived since their introduction, their born status as a Noble or commoner, and their Gender could indicate whether Martin will kill them off over the course of the series.

1.1 Data

Using a previously compiled dataset of 916 characters webscraped from the ASoIaF Wikipedia, data source listed in section 4.2. We set about cleaning and organizing the data, and removing 15 characters that have no introduction chapters or wildly incorrect data, meaning they are only referred to throughout the books, and not 'on stage' at any point.

We created the 'lifetime' number of the characters by adding up the total number of chapters of each book they appear, then subtracting the number of chapters before they were introduced, and if killed, the number of chapters in their last book after they've died. We then created a 'death' indicator, determining whether or not the character is alive by the end of Book 5 of ASoIaF. Though a small number of characters have come back from the dead to some degree, we counted their first 'death' in our Is.Dead indicator variable. The InBook variable is treated in a narrative and not chronological manner, since the texts are often vague with dates and ages.

The Gender and Nobility indicators we took directly from the dataset. For our Gender indicator, we use 1 to denote males, and 0 to denote females. For our Nobility Indicator, we use 1 to denote a noble-born character, and 0 to denote a non-noble-born character. Some of the Nobility indicators are blurred outside of the Westeros feudal structure, so we used the same reasoning as our source. For characters in non-standard medieval structures, Wilding military commanders, and obviously rich and powerful Essoss characters were counted as Noble. Night's Watch characters were determined by their status before joining. Noble bastards were decided on a case-by-case basis.

We then created a matrix of the number of characters listed as dead or still alive, broken down by gender and nobility [Table 1]. Our models were then constructed and simulated in Rjags[7].

2 Model

2.1 Weibull Model

We use the Weibull distribution to model the hazard function, or the probability a character lives for a certain number of chapters. Our Weibull model has two parameters, α and β, which we use to control the model's shape.

$$
\begin{aligned}
&x \sim Weibull(\alpha, \beta) \\
&\pi(x|\alpha, \beta) = (\alpha\beta)x^{\alpha-1}e^{-\beta x^{\alpha}},\ where\ x, \alpha, \beta > 0 \\
&\pi(\beta, \alpha, x^n) \sim G\Big(n+1, \sum x_i^2\Big),\ where\ G(a,b),\ a = shape, b = rate \\
&\pi(\alpha|x^n) = \frac{\alpha_n \Pi x_i^{\alpha-1} n!}{(\sum x_i^{\alpha})^{n+1}}
\end{aligned}
$$

Assume non-informative prior for (α, β), where $\alpha > 0$ and $\beta > 0$. We are interested in the median chapters of survival θ and the probability of characters survive at least k books after their introduction, denoted as λ_k. The probability of a living character's death within another 70 chapters, denoted as $D(t + t_0|t_0)$, is another statistic of interest for our model. We use 70 chapters because the average number of chapters in an ASoIaF book is 68.8, and we are not yet sure how many chapters the sixth book, The Winds of Winter[10], will have upon release.

$$\theta = \Big(\frac{log2}{\beta}\Big)^{\frac{1}{\alpha}}$$
$$\lambda_k = e^{-\beta(70k)^{\alpha}} \; where \; k = 1, 2, 3, 4, 5$$
$$D(70 + c_0|c_0) = 1 - e^{\beta(t_0)^{\alpha} - \beta(70+c_0)^{\alpha}}, \; where \; 0 < c_0 < 345$$

2.2 Generalized Linear Model

We use this model to analyze the odds ratio of characters death across the entirety of ASoIaF for gender controlling for nobility, and the odds ratio for nobility controlling for gender. Let $X_G = 1$ if the gender is male and $X_G = 0$ if the gender is female. $X_N = 1$ if the character is nobility and $X_N = 0$ if not. $Y = 1$ if the character is dead, and $Y = 0$ otherwise. And $i \in \{1, 2, ..., N\}$, where N is the total number of characters in the first five books of ASoIaF. We use the alive or dead matrix we constructed from our data in [Figure 1] for our data. The likelihood is:

$$log(\frac{P(y_i = 1|x_{iG}, x_{iN})}{1 - P(y_i = 1|x_{iG}, x_{iN})}) = c + \beta_G x_{iG} + \beta_N x_{iN}$$

Controlling for gender or nobility, we can obtain the parameter of nobility β_N or gender β_G. Here, denote $x_{ij} = 1$ as the x_{ij_1}

$$log(\frac{P(y_i = 1|x_{ij_1})}{1 - P(y_i = 1|x_{ij_1})}) - log(\frac{P(y_i = 1|x_{ij_0})}{1 - P(y_i = 1|x_{ij_0})} = \beta_j, \; j = N, G$$

2.3 Hierarchical Generalized Linear Model

We constructed a generalized linear mixed model. First, we split the data to four groups according to gender and nobility and got the cross table for Books 1-5 as [Table 1].

Here, we consider the whole data as population and conduct a double randomize sampling. First, randomly select n_{ij}, and then randomly sample from group j,

which is
$$n_{ij} \sim Unif(p_1 N_j, p_2 N_j), \; j = \{1, 2, 3, 4\}$$

Here, N_j is the total number of case in group j and $0 < p_1 < p_2 < 1$. We use p_i to denote the probability of death for character i, x_{iG} denote the gender type and x_{iN} denote the nobility. The random effect logit model is stated as follow:

$$
\begin{aligned}
&[Y_i|p_i] \sim Bin(p_i, n_i)\\
&logit = \beta_i\\
&[\beta_i|\mu_i, \sigma^2] \sim N(\mu_i, \sigma^2)\\
&\mu_i = \alpha_0 + \alpha_N x_{iN} + \alpha_G x_{iG} + \alpha_{NG} x_{iN} x_{iG}
\end{aligned}
$$

where

- α_0: intercept
- α_N: main effect due to Nobility
- α_G: main effect due to Gender
- α_{NG}: interaction between Gender and Nobility

which implies:

$$logit(p_i) = \alpha_0 + \alpha_N x_{iN} + \alpha_G x_{iG} + \alpha_{NG} x_{iN} x_{iG} + \epsilon_i$$

where $\epsilon_i = \beta_i - \mu_i \sim N(0, \sigma^2)$. Thus, ϵ_i is a random effect that account for the overdispersion in the data.

Here we use independent and proper prior as follows:

$$
\begin{aligned}
&\alpha_i \sim N(0, 100000),\ i = \{0, N, G, NG\}\\
&\sigma^{-2} \sim G(0.001, 0.001)
\end{aligned}
$$

3 Data Analysis

All the results and test for the three models are shown in [Table 2]. To check our models' convergence, we use the Gelman-Rubin Convergence Diagnostic, which shows that as our number of iterations approaches a large number, the ratios of our model parameters converge to 1. If the ratios converge to 1, it means that the bias of the estimated chains has approached zero, which means that our simulation's parameters are stable.

3.1 Weibull

After simulations, we found that our data has approximately a $Weibull(\alpha, \beta)$ distribution. Our sample from the posterior distribution derived in the [section 2.1]. α has a mean of 0.954, and a 95% credible interval of $(0.904, 1.006)$, and β has a mean of 0.017, and a 95% credible interval of $(0.013, 0.021)$.

Within our Weibull model, we look at λ_k, which is the simulated probability of characters who survive $70 \times k$ chapters, or around k ASoIaF books. We also construct θ, which is the median number of chapters that characters have survived. For θ, the mean is 49.362, and the 95% credible interval is $(45.439, 53.458)$, which suggests that the median survival time is approximately 49. As for λ_k, the means are $(0.380, 0.153, 0.063, 0.0265, 0.0113)$ for $k = (1, 2, 3, 4, 5)$ respectively.

This indicates that characters' chances of survival start at one third for a book's length of chapters(70), and are cut by more than half for each successive length of 70 chapters. We then observed the simulated chances of individual Point of View characters living another 70 chapters.

We observe from our results for our Weibull distribution of the survival chances for characters in books 1-5 in [Figure 1], with Credible Intervals for the variables included.

3.2 Generalized Linear Model

After we implement the model in R, the results show that β_N has a mean of 0.675, and a 95% credible interval of $(0.265, 1.097)$. The posterior density plot for parameters is in [Figure 2] The posterior median is $exp(\beta_N) \approx 0.536$. Thus, controlling for the Gender, the odds of a noble-born character being dead are 0.536 times of a non-noble-born one. For β_G, the mean is -0.625, the 95% credible interval is $(-0.914, -0.342)$. The posterior median is $exp(\beta_G) \approx 1.960$. Therefore, controlling for the Nobility, the odds of a male character being dead are 1.960 times of a female one.

Observing from the model, Gender and Nobility are significant predictors for the death rates of characters in the series. Gender has a $\beta_G > 0$, indicating that female characters are more likely to survive through the first five books. And Nobility has a $\beta_N < 0$, indicating that Noble characters are more likely to survive through five books' worth of chapters than their commoner counterparts of the same gender.

3.3 Hierarchical Generalized Linear Model

We run the model in R and get the results (see Figure 3) that α_G has a mean of 0.632, and the 95% credible interval is $(0.364, 0.891)$. The posterior median is $exp(\alpha_G) \approx 1.895$. Therefore, controlling for the Nobility, the odds of a noble-born character being dead are 1.895 times of a non-noble-born one.

α_N has a mean of -0.708, and the 95% credible interval is $(-0.909, -0.486)$. The posterior median is $exp(\alpha_N) \approx 0.493$. Therefore, controlling for the Gender, the odds of a male character being dead are 0.493 times of a female one.

And α_{NG} has a mean of -0.198, and the 95% credible interval is $(-0.494, 0.120)$. The posterior median is $exp(\alpha_{NG}) \approx 0.848$. Since 0 is included in the 95% CI, this interaction between Gender and Nobility doesn't seem work very well.

4 Discussion

4.1 Convergence Diagnostics

We use Gelman and Rubin Convergence Diagnostic for all the three models, and draw the plot of HGLM in [Figure 4]. We also append the results of potential scale reduction factors in [Table 2]. Based on the values, no parameters are substantially larger than 1, which indicates our models are all convergent and our observed parameters are stable.

4.2 Models

With our Weibull Model, we find that the odds of survival drop precipitously as the number of chapters increases, and the median value of chapters is around 50, which is less than the length of one ASoIaF book. This makes sense, because a large number of characters die, and a subset of those characters die within the same book they're introduced, ([48, 54, 55, 9, 26] for books [1,2,3,4,5] respectively.) We note that since the Credible Interval of the Alpha parameter includes 1, with an estimate of 0.954, our model is plausibly an Exponential distribution [$\beta e^{-\beta * x}$], so the characters' survival chances could be memoryless. When attempting point estimates for the survival chances of single characters, we noted that each Point-of-View character had a roughly 60-70 percent chance of death no matter how many chapters they had previously survived. Five example Point-of-View characters, with the number of chapters they've lived thus far, their probabilities of survival means and Credible Intervals are included in [Table 2] to demonstrate.

We were unable to find a convergent or conclusive model that included an indicator variable for the number of chapters a character has lived. Median, Mean, and Interquartile Range indicator variables describing the number of chapters each character had were unsuccessful in modeling the data within a Hierarchical Generalized Linear Model. And while we were unable to also merge the Weibull and Generalized Linear Models in a way that allowed for predictions given Chapters, Gender, and Nobility, we feel that our models can help predict a character's chances of survival in the series. We found that Nobles and Women were more likely to survive ASoIaF thus far. This would make sense, since women in the series generally do not participate in wars, and the nobility are shielded from most of the collateral damage of battles and hardship.

Since ASoIaF is a work of fiction, who lives and dies is up to George R.R. Martin, but there appears to be strong correlations between a character's Gender, born social status, and their chances of survival within the series, as well as a memoryless survival rate in terms of chapters featured.

References

Source for our dataset. http://allendowney.blogspot.com/2015/03/bayesian-survival-analysis-for-game-of.html

[1] George R. R. Martin, A Game of Thrones, Bantam, 1996.

[2] George R. R. Martin, A Clash of Kings, Bantam, 1999.

[3] George R. R. Martin, A Storm of Swords, Bantam, 2000.

[4] George R. R. Martin, A Feast for Crows, Bantam Spectra, 2005.

[5] George R. R. Martin, A Dance with Dragons, Bantam Spectra, 2011.

[6] R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[7] Plummer, M., Stukalov, A., Denwood, M., Plummer, M. M. (2015). Package rjags.

[8] Vale, R, (2014). Bayesian Prediction for The Winds of Winter, URL https://arxiv.org/abs/1409.5830

[9] Downey, A, (2015). Bayesian Survival Analysis for "Game of Thrones", URL http://allendowney.blogspot.com/2015/03/bayesian-survival-analysis-for-game-of.html

[10] George R. R. Martin, The Winds of Winter, Bantam Spectra, expected 2018.

Table 1: Cross classified

Nobility	Gender	Death	Alive	Total
Yes	Male	97	240	337
Yes	Female	9	75	84
No	Male	163	244	407
No	Female	24	49	73

Table 2: Results

		Point Esitimate			Gelman Rubin test		
		Estimate	2.5%	97.5%	Point est	Upper CI	Psrf
Weibull	α	0.954	0.904	1.006	1.000	1.000	1.000
	β	0.017	0.013	0.021	1.000	1.000	
	$\lambda 1$	0.38008	0.354968	0.40558	1.000	1.000	
	$\lambda 2$	0.15343	0.134951	0.17312	1.000	1.000	
	$\lambda 3$	0.06335	0.051350	0.07674	1.000	1.000	
	$\lambda 4$	0.02657	0.019426	0.03498	1.000	1.000	
	$\lambda 5$	0.01129	0.007388	0.01622	1.000	1.000	
	θ	49.362	45.439	53.392	1.000	1.000	
	$JonCon(68)$	0.66316	0.600899	0.71746	1.000	1.000	
	$Arianne(117)$	0.65660	0.601321	0.70450	1.000	1.000	
	$Asha(178)$	0.65098	0.601374	0.69361	1.000	1.000	
	$Davos(225)$	0.64768	0.601359	0.68737	1.000	1.000	
	$JSnow(343)$	0.64149	0.601204	0.67595	1.000	1.000	
GLM	c	-1.030	-1.442	-0.626	1.000	1.010	1.000
	β_G	0.675	0.265	1.097	1.000	1.010	
	β_N	-0.625	-0.914	-0.342	1.000	1.000	
HGLM	α_0	1.082	0.923	1.253	1.000	1.020	1.120
	α_N	0.632	0.364	0.891	1.060	1.160	
	α_G	0.708	-0.909	-0.486	1.010	1.040	
	α_{NG}	0.161	-0.518	0.194	1.04	1.05	

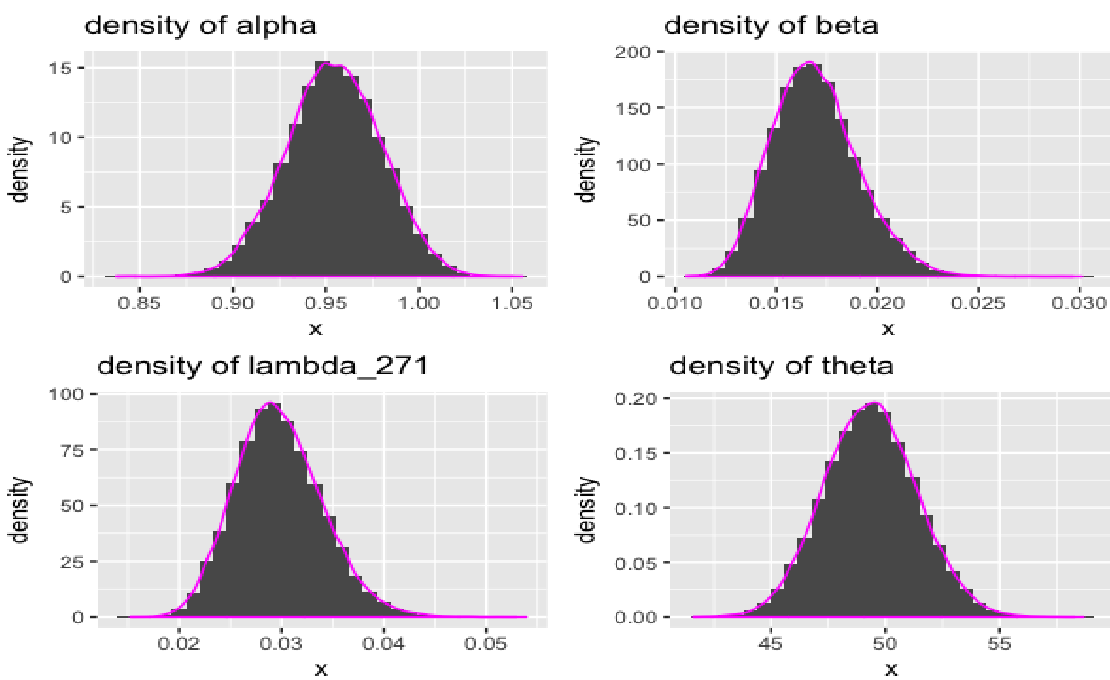
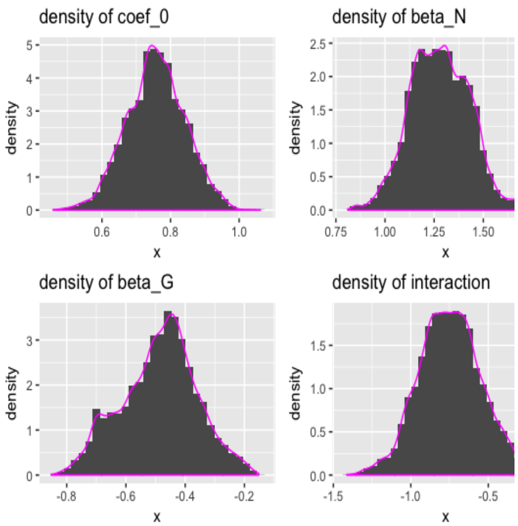
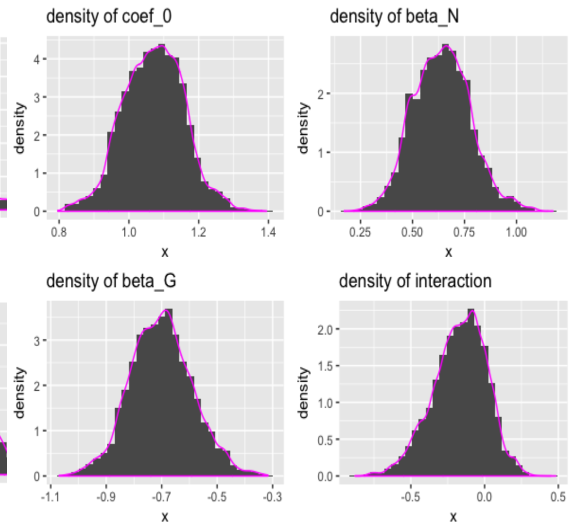


Figure 1: Weibull

Figure 2: Death for book 1-5



(a) Random range 20-40



(b) Random range 5-70

Figure 3: Random Effect logit model

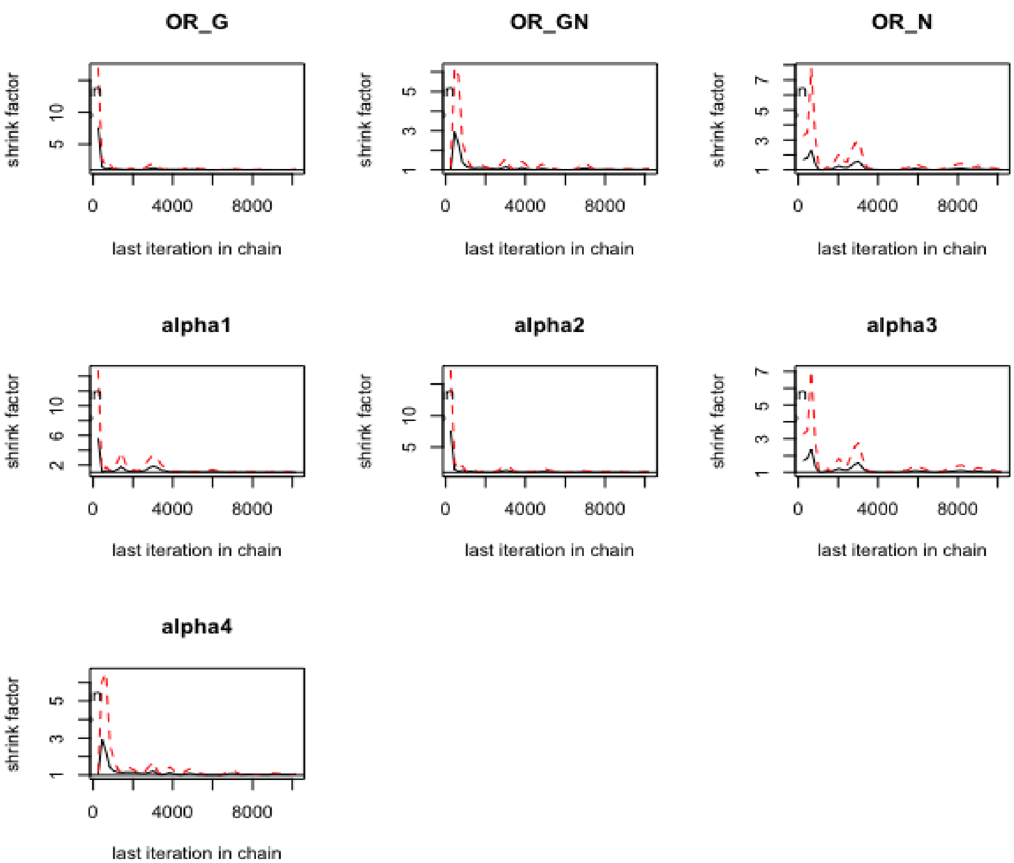


Figure 4: HGLM Gelman Plot