Published in final edited form as:

Stat Med. 2013 October 15; 32(23): 4044–4056. doi:10.1002/sim.5820.

Beta Prime Regression with Application to Risky Behavior Frequency Screening

Alexander Tulupyev¹, Alena Suvorova¹, Jennifer Sousa², and Daniel Zelterman^{2,*}

¹Saint Petersburg State University, Faculty of Mathematics and Mechanics and Saint Petersburg Institute for Informatics and Automation, Russian Academy of Sciences, Russia

²Department of Biostatistics, School of Epidemiology and Public Health, Yale University, New Haven, CT 06520

Abstract

Our aim is to model the frequency of certain behavioral acts, especially those that are likely to transmit communicable diseases between persons. We develop a generalized linear model based on the beta prime distribution to model the responses to a survey question of the form, “When was the last time that you engaged in this behavior?” Intuitively, individuals reporting more recent events are more likely to have greater frequency of the risky behavior. The beta prime distribution is especially suited to this application because of its long tail. We adjust for length-biased sampling. We show how to use this distribution as the basis of a linear regression model that accounts for differences in demographic and psychological characteristics of the respondents. We discuss estimation of parameters, residuals, tests for heterogeneity of these parameters, and jackknife measures of influence. The methods are applied to a survey of alcohol abuse use among individuals who are at high risk for spreading HIV and other communicable diseases in a study conducted in St. Petersburg, Russia.

Keywords

length bias; HIV infection; regression diagnostics; parameter heterogeneity; recall bias

1 Introduction

We want to develop a measure of individual risk indices for transmission of contagious diseases such as HIV that are based on interviews of individuals with known risky behaviors. These behaviors will typically include intravenous drug use, substance abuse, and unprotected sex with strangers. Specifically, we need an estimate of the frequency of these risky behaviors. The population is well known to be heterogeneous even after taking the effects of covariates into account. The survey that motivates this work is described here, before we go further into the analysis.

In 2006–8, the St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), in collaboration with the St. Petersburg Municipal AIDS

---

*corresponding author: daniel.zelterman@yale.edu.

Center, conducted three closely-related pilot studies on HIV-positive people in terms of their risky behavior, their adaptive styles, other psychological traits, and socio-demographic characteristics. The major goal of these studies was to identify patterns of maladaptive behaviors of HIV+ people that could lead to higher rates of HIV transmission. Their psychological defense mechanisms and coping strategy expression levels were assessed along with correlates of their tendencies towards risky behaviors such as unprotected sex, inclination to risk, sensation seeking, drug and alcohol abuse, and deviations from the standard highly-active antiretroviral therapy (HAART) regiment. The list of questions and selected preliminary results have appeared [1–4]. These studies provided motivation for the present work.

We concentrate on the data on alcohol abuse extracted from the pilot studies. In 2007 there were 306 HIV+ study participants. We choose to examine only those 252 with complete data, thereby facilitating the fitting of regression models in Section 4. Among those subjects analyzed there were 112 (44%) males and 140 (56%) females. Most of the participants were 35 years old or younger (70% of males and 75% of females). A marginal analysis of the time of last abuse of alcohol is described at the conclusion of Section 2. A histogram of these values appears in Fig. 1, illustrating the extremely long tail of the data, and motivating the use of the beta prime sampling model used in this study. A more complete regression analysis incorporating covariate values is summarized in Section 4.

The covariates we examined included the sex of the respondents and their age in years. Four additional composite measures of psychological well-being were summarized as: coping with stressful experiences by emphasizing personal growth; sensation seeking and welcoming new experiences; defense mechanisms; and the tendency to engage in risky behaviors. These are referred to as: Coping; New; Defense; and Risky, respectively.

Standard surveys of substance abuse frequency ask the participants how often they engaged in the risky behavior in the previous 30 days (SAMHSA, [5], for example). In contrast, our survey asked participants: “When did you last participate in this behavior?” We argue, on intuitive grounds, that a more recently-reported event is indicative of more frequent engagement by this individual. This question asks the respondent to refer to the most recent event and, as such, minimizes the amount of bias due to recall error that is more likely when longer time intervals are involved. (Over 30% of all reported values were more than 30 days prior to the interview.) In addition, some respondents answered with a vague estimate such as “about a year ago.” These values had to be recoded with an approximate numerical value. We will show that such approximations have little influence on our fitted models.

Let us denote by t , the reported time since last abuse of alcohol. Values of $t = 0$ were recorded for individuals who were either under the influence of alcohol at the time of their interview or else responded that their last abuse of alcohol was within a day of their interview. This convention proved to be a problem in the likelihood function and we discuss the choice of coding these values on regression parameters in Section 4.

The screen in a survey of this type is a random event in the life of the participant, so the last-event question is more likely to be asked in the middle of a larger, rather than shorter, time

interval between events. This screen then has an inherent *length bias* favoring longer intervals. Screens, in general, and for breast cancer, as an example, are similarly biased towards detecting slow-growing, pre-clinical tumors rather than finding the faster growing cancers that pose the greater risk [6–8]. In Section 2 we demonstrate that the functional form of the beta prime sampling model facilitates the adjustment for length bias.

In Section 2 we approximate the distribution of the last-event times as a scale member of the beta prime family. The *beta prime* is a family of distributions with density functions of the form

$$f(t|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} t^{\beta-1} / (1+t)^{\alpha+\beta} \quad (1)$$

defined on $t > 0$ for parameters $\alpha > 0$ and $\beta > 0$.

The density function (1) is sometimes referred to as the *beta distribution of the second kind*. It is a member of the Pearson Type VI family and can also be obtained as a multiple of the *F*-distribution. The distribution of $t=(1+t)$ is that of the usual beta distribution defined on $(0-1)$. See Johnson, Kotz, and Balakrishnan ([9] p 248). The “beta prime distribution” entry in *Wikipedia* [10] is an accessible and comprehensive reference. This reference also describes four parameter generalizations of the beta prime distribution.

The review by McDonald and Xub [11] illustrates relationships between this distribution and many other generalizations of the beta distribution. Others have studied this distribution in a variety of theoretical settings [12–15]. None of these references examine the beta-prime distribution as the basis of a regression model.

The following two statistical tools will be useful in developing diagnostics for regression models in Section 3. An important diagnostic measure we will employ is based on the jack-knife. Suppose observations $\mathbf{x}' = (x_1, x_2, \dots, x_n)$ are independently sampled from a (more general) distribution with density function $f(x|\theta)$ and real-valued parameter θ . If the i -th observation x_i is deleted (jackknifed) and the model is refitted, then the approximate change in the maximum likelihood estimate is

$$D_i(\theta) = \frac{\partial / \partial \theta \log f(x_i|\theta)}{-\sum_j \{ \partial / \partial \theta \}^2 \log f(x_j|\theta)} \quad (2)$$

evaluated at $\theta = \hat{\theta}$ the maximum likelihood estimator of θ based on the full dataset. This statistic is sometimes called the *dfbeta*. See also Pregibon [16] or Adewale and Xu [17]. Recognize the denominator of (2) as the observed information for θ and this sum is the same for every value of $i = 1, \dots, n$.

We will also make use of the test for overdispersion developed by Zelterman and Chen [18]. The test for overdispersion rejects the null hypothesis of a constant value of θ against an alternative in which θ varies between observed values of x_i for large values of the statistic

$$U(\mathbf{x}|\theta) = \sum_i U_i(x_i|\theta) = \sum_i (\partial/\partial\theta)^2 \log f(x_i|\theta) + \{(\partial/\partial\theta) \log f(x_i|\theta)\}^2 \quad (3)$$

where θ is evaluated at $\hat{\theta}$.

Briefly, U is the difference between two estimates of the observed information of θ . If the observed information is too small, then there is evidence of overdispersion in the parameter. In words, $U(\mathbf{x}|\hat{\theta})$ is large if the observed log-likelihood is “flatter” than anticipated by the distribution. Under the null hypothesis of a homogeneous parameter value, U is approximately normally distributed with zero mean. The diagnostic measure we will use is the i -th term in the sum, namely, $U_i = U(x_i|\hat{\theta})$ as a measure of influence that the i -th observation has on this test of overdispersion.

Informally, large absolute values of D_i indicate an observation's influence on the fitted estimate of the parameter. Large values of U_i show that this observation inflates the estimated standard error of the parameter's estimated value.

In Section 2, we derive the beta prime distribution as a gamma mixture of exponential distributions and correct the resulting distribution for length bias. We show that maximum likelihood estimates exist in the marginal distribution for any set of observed last-event values. In Section 3, we develop regression models for the beta prime distribution including residuals and diagnostics for influence and overdispersion. In Section 4, we apply these methods to our motivating example involving frequency of alcohol abuse among HIV+ persons in St. Petersburg, Russia.

2 Sampling Distribution and Marginal Data Analysis

We begin by assuming that the time T between successive events follows an exponential distribution with density function

$$f_0(t|\lambda) = \lambda e^{-\lambda t}$$

for $t > 0$, conditional on the rate parameter $\lambda > 0$.

We also assume that the rate λ is heterogeneous across the population of individual event times. For mathematical convenience, we found it useful to assume that λ follows a gamma distribution with density function

$$g(\lambda) = g(\lambda|\theta, \sigma) = \sigma^\theta \lambda^{\theta-1} e^{-\sigma\lambda} / \Gamma(\theta)$$

for $\theta > 0$ and $\sigma > 0$.

The marginal distribution of inter-event times then has the density function

$$f_1(t) = f_1(t|\theta, \sigma) = \int f_0(t|\lambda) g(\lambda) d\lambda = \theta/\sigma (1+t/\sigma)^{\theta+1}. \quad (4)$$

This is a scale-parameter member of the beta prime family of distributions (1).

Screening studies are subject to length bias, whereby the probability that the screening event samples a given time interval is itself proportional to the length of the time interval being sampled. The density function of the distribution that corrects for this length bias in (4) is proportional to $tf_1(t)$. This yields the sampling distribution with density function

$$f(t|\alpha, \sigma) = \{\alpha(\alpha+1)/\sigma\} \frac{t/\sigma}{(1+t/\sigma)^{\alpha+2}} \quad (5)$$

that we will concentrate on in the remainder of this narrative. The parameter θ in (4) is equal to $\alpha+1$ in (5).

This distribution is also a member of the beta prime family (1). We can see that $\sigma > 0$ in (5) represents a scale parameter and $\alpha > 0$ controls the shape of this distribution. Mathematical properties of this distribution are also examined in Zelterman, Tulupyev *et al.* [19]. Several examples of this density function are displayed in Fig. 2 for different values of α and σ in the range of values encountered among the fitted models of Section 4. The data displayed in Fig. 1 are clearly sampled from a very long-tailed distribution. The long, polynomial tails of (5) are especially suited for these data.

The cumulative distribution corresponding to (5) is

$$F(t|\alpha, \sigma) = 1 + \alpha(1+t/\sigma)^{-\alpha-1} - (\alpha+1)(1+t/\sigma)^{-\alpha}.$$

The k -th moment in (5) is given by

$$ET^k = \int t^k f(t) dt = \sigma^k \Gamma(k+2) \Gamma(\alpha-k) / \Gamma(\alpha)$$

for $k < \alpha$ from which we compute the mean

$$ET = 2\sigma / (\alpha-1) \quad (6)$$

which defined for $\alpha > 1$ and the variance

$$\text{Var } T = 2\sigma^2(\alpha+1) / \{(\alpha-1)^2(\alpha-2)\}$$

which is defined for $\alpha > 2$.

In (6) we see that smaller values of σ and/or larger values of α result in smaller expected values of T . Similarly, these conditions correspond to estimates of more frequent episodes of the behavior that we are modeling.

In the fitted models of Section 4, we typically estimated parameters outside of the range where the mean of the fitted distribution is undefined. For this reason we summarize the fitted model in terms of its median and other quantiles in Fig. 3.

Parameter estimation using maximum likelihood is valid as a result of Lemma 1, in the Appendix. Specifically, for any set of observed times $t_i > 0$ there exist critical values of the likelihood satisfying $\hat{\alpha} > 0$ and $\hat{\sigma} > 0$. Mathematical expressions for the expected information are also given in the Appendix, but the variance estimates for regression parameters are more easily calculated numerically in **R**.

To complete this section, consider a univariate, marginal analysis of the single question concerning the time of the last abuse of alcohol. More detailed data analysis and regression models are examined in Section 4.

The likelihood at (5) is identically zero at the observed value $t = 0$ and these were recoded as $t = .5$. The effect of this recoding on the inference is discussed again in Section 4. The marginal, sample average time since last abuse of alcohol is 112.2 days with a median of 7, consistent with the highly-skewed nature of the $N = 252$ responses analyzed, illustrated in the histogram of Fig 1.

The maximum likelihood parameter estimates for the beta prime distribution in (5) are $\hat{\alpha} = .50$ ($SE = .04$) and $\hat{\sigma} = 1.08$ ($SE = 0.18$). These estimates and their standard errors were obtained numerically using the **nlm** routine in **R**. The median of the fitted model is 3.2 days with quartile range (0.82, 15.6). The mean of this fitted distribution is undefined.

3 Regression Models and Diagnostics

Regression for the shape α and scale σ parameters in (5) will be described next. Suppose that the i -th respondent has an observed last-event time $t_i > 0$. A vector of covariate or explanatory values $\mathbf{x}_i' = (x_{i1}, x_{i2}, \dots, x_{ip})$ is also associated with this individual. We assume that conditional on $\mathbf{x}_i$, the last-event time t_i for the i -th subject is sampled from a beta prime distribution $f(t_i/\alpha_i, \sigma)$ where the α_i follow the model

$$\alpha_i = \exp(\boldsymbol{\beta}' \mathbf{x}_i) \quad (7)$$

for a vector of regression coefficients $\boldsymbol{\beta}' = (\beta_1, \beta_2, \dots, \beta_p)$ that will be estimated using maximum likelihood. The link function in (7) between the shape parameter α_i and the linear predictor $\boldsymbol{\beta}' \mathbf{x}_i$ assures that $\alpha_i > 0$ for every possible value of $\boldsymbol{\beta}$ and $\mathbf{x}_i$.

We also fitted models in which the scale parameter σ varies according to subject-level covariates. Specifically, the model (7) was combined with regression models in which

$$\sigma_i = \exp(\gamma' x_i) \quad (8)$$

were fitted for a vector of regression coefficients γ .

We obtained maximum likelihood estimates of the regression coefficients β and γ in $\mathbf{R}$ using the **nlm** routine. This routine also has a useful numerical approximation to the Hessian matrix that we used to approximate the variances of these estimates $(\hat{\beta}, \hat{\gamma})$. This was the method used to obtain the maximum likelihood estimated parameter values and their approximate standard errors in Tables 1, 2, and 3. Statistical significance of these regression coefficients was obtained by approximating the distribution of their maximum likelihood estimates using a normal distribution.

Table 1 examines model (7) for all covariates and a single σ for all subjects, and Table 2 summarizes fitted examples of model (8) and a single value of α for all subjects. The maximum likelihood estimates for β and γ are given in Table 3 for models in which both α and σ vary with covariate values. More details and interpretations are described in Section 4.

The beta prime distribution exhibits long tails and the expected value of T is not defined in any of the fitted distributions we obtained. Residuals cannot be defined in the usual manner of the differences between observed and expected values. Figure 3 displays a plot of the observed values along with robust estimated and smoothed percentiles of the fitted distribution. The robust estimates minimize the absolute deviations about the median, on a log scale. That is,

$$(\hat{\beta}, \hat{\sigma})_{\text{robust}} = \arg \min_{\beta, \sigma} \sum_i \left| \log(t_i) - \log F^{-1}\{1/2 | \alpha(\beta' x_i), \sigma\} \right|.$$

The specific model fit illustrated in Fig. 3 uses Coping in α and one σ for all subjects.

In the sample of size 252, we would expect to see 12.6 observations above the 95-th percentiles and 11 were observed so that the upper tail of the fitted model agrees well with the empirical data. Only one observed value fell below the fitted 5-th percentiles. This lower tail behavior is partly explained by the way we coded the seven observed values of $t_i = 0$ in the analysis. The topic of this recording is discussed again in the following section.

The regression models were fitted using traditional maximum likelihood because of the increased sensitivity and ease of estimating standard errors from the numerical approximation of the Hessian matrix provided by **nlm**. We next illustrate diagnostics for this model based on similar measures for other generalized linear models.

The jackknife diagnostics in (7) for each β_j is

$$D_i(\beta_j) = x_{ij} \{ (2\alpha_i + 1) / (\alpha_i + 1) - \alpha_i \log(1 + t_i / \sigma_i) \}$$

divided by the observed information for β_j and the corresponding jackknife diagnostic for γ_j is

$$D_i(\gamma_j) = x_{ij} \{(\alpha_i + 2)t_i / (\sigma_i + t_i) - 2\}$$

divided by the observed information for γ_j .

The overdispersion diagnostics in (3) corresponding to β_j are

$$U_i(\beta_j) = x_{ij}^2 \left[\frac{4\alpha_i + 1}{\alpha_i + 1} - \frac{\alpha_i(3\alpha_i + 1)}{\alpha_i + 1} \log(1 + t_i / \sigma_i) + \alpha_i^2 \{\log(1 + t_i / \sigma_i)\}^2 \right]$$

and

$$U_i(\gamma_j) = x_{ij}^2 \left(\alpha_i^2 t_i^2 + 4\sigma_i^2 - 5\alpha_i t_i \sigma_i - 2\sigma_i t_i \right) / (\sigma_i + t_i)^2$$

is the diagnostic for γ_j .

The D_i and U_i diagnostics are examined in the following section with our analysis of the motivating dataset.

4 Application

We found that a useful modeling strategy was to identify useful covariates in separate models for α and σ and then combine these in joint models for both of these two parameters. These models are summarized in Tables 1, 2, and 3, respectively. The density function at (5) is identically zero at $t = 0$ and at these values, the log-likelihood is undefined. Zero reported last time of abuse of alcohol were replaced by 0.5. The effects of this recoding will be discussed again at the end of this section.

Table 1 presents the maximum likelihood estimated marginal regression coefficients for each of six explanatory covariates (Age, Sex, Coping, New, Defense, and Risky) in model (7) for α , along with the full regression model containing all six of these covariates. These models assume that the shape parameter α varies across subjects and the beta prime distribution has a fixed value of the scale σ for all subjects.

Only two covariates (Age and Coping) in Table 1 were statistically significant in marginal models. A fitted model containing only these two are also included in this table but Age loses its statistical significance in the presence of Coping. We conclude that Coping is important in determining differences in the shape parameter α . The negative regression coefficient is interpreted to mean that higher levels of Coping scores result in smaller values of α and corresponding longer times since the last abuse of alcohol. That is, individuals who are better able to cope with stressful situations are less frequent abusers of alcohol.

A similar analysis held the shape parameter α fixed and modeled the scale σ in terms of all covariates, both marginally and jointly. These models fitted using maximum likelihood and are summarized in Table 2. Age, Coping, New, and Risky are all statistically significant in marginal models. Taken together, Risky is not statistically significant in the model for σ in (8). The effect of Age has modest significance level of .096 in this model with all four explanatory variables and this covariate will be examined further.

Table 3 we summarized all combinations of covariate models in which Coping is used to model the shape α and Age, Coping, and New, individually, pair-wise, and jointly are used in a model for the scale σ . The final regression model we chose among all of these was

$$\begin{aligned}\alpha &= \exp(\beta_0 + \beta_1 \text{ Coping}) \\ \sigma &= \exp(\gamma_0 + \gamma_1 \text{ New})\end{aligned}\quad (9)$$

with maximum likelihood fitted parameter values indicated in Table 3.

In this fitted model both β_1 and γ_1 exhibit extreme statistical significance levels of .006 and .003, respectively. Model (9) has a log-likelihood of 1151.74 with four parameters. In contrast, this fit is not very different from the likelihood value of 1147.36 for the 14 parameter model containing all possible regression coefficients for both α and σ .

The estimated $\hat{\beta}_1$ and $\hat{\gamma}_1$ are both negative indicating that larger Coping and smaller New scores are associated with lower frequency of alcohol abuse. These are individuals who are better at coping with stressful situations and/or less likely to welcome new sensation experiences. The correlation between New and Coping was negligible (.014) so individuals with either of these characteristics were neither more nor less likely to exhibit the other.

Regression diagnostics for model (9) are given in the panel display of Fig. 3. In each of these diagnostics, we have identified the observations with last-event times $t_i = 0$ and $t_i = 1$ as open circles. One individual claiming ten years since his last abuse of alcohol is displayed with a star. The first panel plots the robust estimated median times against the observed last-event times, both on a log scale. At any given estimated median time, we can also obtain the estimated quantiles of the fitted beta prime distribution. These will vary slightly for observations with the same estimated median and so have been smoothed using *loess* to provide a better idea about the typical fitted distribution. From left to right, estimated median times increase indicating reduced estimated frequency of alcohol abuse.

The jackknife change in regression slopes D_i indicate that the smallest recorded last alcohol abuse times were influential in estimating the slope of New in the scale σ and, to a smaller degree, the slope of Coping in α . Intuitively, many short last-event times should result in smaller estimates of the scale parameter. On the other hand, the low influence of the extremely long time demonstrates useful attributes of the beta prime as a sampling model.

The plot of overdispersion measures U_i show that the smallest recorded last abuse times are also influential in inflating the approximated variance of the estimated slope $\hat{\beta}_1$ of Coping in

α but not that of the slope γ_1 of New in σ . The single extreme observation denoted with a star in this plot corresponds to one individual who claimed not to have had an alcoholic drink in ten years. The validity of such an observation is questionable, but the conclusion of this figure is that such an observation neither results in a great change in the estimated value of either regression coefficient nor in an inflation in the estimated standard errors of the estimated regression slopes.

Having identified the smallest last-event times as being influential, we conducted a small study to see how the numerical coding of these values influences the maximum likelihood estimated regression slopes. In all of the analysis of this section up to this point, we coded last time of alcohol abuse of “yesterday” responses as 1 and any response less than yesterday as .5. In this final examination, we considered different codings for these two responses and refitted the regression model (9) for each.

Fig. 4 summarizes the percent relative change in each of the four estimated regression coefficients when we recode these last abuse times. Times for less than one day are given and times for yesterday are coded as twice these values. Values at .5 refer to the original coding and represent no change. The estimated coefficient for the intercept β_0 in $\alpha(\beta)$ changes a great deal, but the slope β_1 for New in α at (9) has a negligible change. The changes in the γ_1 slopes and intercept γ_0 in the scale parameter σ are small and appear to offset each other by the same amount.

5 Conclusions

The benefit of asking about the time since the “last event” rather than the number of activities in the last 30 days minimizes the chance of recall bias and has a direct link with the frequency of these activities.

The long tail of the data motivates the use of an unconventional sampling distribution. The beta prime model has a long, polynomial tail making it suitable for modeling such data. The functional form facilitates an adjustment for length biased sampling, further adding to its utility. The long-tailed beta prime distribution is neither greatly influenced by the presence of extremely large observed responses, nor by reasonable choices of coding for extremely short last times of abuse.

In the data analysis, we conclude that the New and Coping measures have high explanatory value in describing the risky behavior involving the last reported abuse of alcohol. Individuals who are better at coping with stressful situations and are less likely to engage in sensation-seeking activities are more likely to be associated with a lower frequency of alcohol abuse.

Acknowledgments

This research was supported by grants P30 MH62294 awarded to the Yale Center for Interdisciplinary Research in AIDS (AT, AS, DZ) and training grant T32 MH014235 support for JS. We are grateful to Elizabeth Nichols for editorial assistance.

References

1. Tulupyeva TV, Tulupyev AL, Paschenko AE. Psychological Traits and Behavior Admissibility. SPIIRAS, SPb. 2006:32. (In Russian).
2. Tulupyev AL, Tulupyeva TV, Paschenko AE, Syvorova AV. An approach to comparison of threatening behavior parameters between social groups based upon incomplete and imprecise data. SPIIRAS Proceedings. 2009; 9:252–261. (In Russian). Available at: <http://mi.mathnet.ru/eng/trspy/v9/p252>.
3. Tulupyeva TV, Tulupyeva AL, Stolyarova EV, Pashchenko AE. An analysis of HIV-positive persons' risky behavior in their adaptive style models (based on interviews obtained from the St. Petersburg Aids-Center patients). SPIIRAS Proceedings. 2007; 5:117–150. (In Russian.). Available at: <http://mi.mathnet.ru/eng/trspy/v5/p117>.
4. Paschenko AE, Tulupyev AL, Tulupyeva TV. Respondents Behaviour Intensity Estimation under the Information Defficiency. SPIIRAS Proceedings. 2008; 7:239–254. (In Russian.). Available at: <http://mi.mathnet.ru/eng/trspy/v7/p239>.
5. SAMHSA (Substance Abuse and Mental Health Services Administration). Performance Measurement/GPRA Tools. 2006. Available at: <http://www.samhsa.gov/Grants/CSAP-GPRA/index.aspx>
6. Zelen M. Optimal scheduling of examinations for the early detection of disease. Biometrika. 1993; 80:279–293. Corr: 1996;83:249.
7. Zelen M. Forward and backward recurrence times and length biased sampling: Age specific models. Lifetime Data Analysis. 2004; 10:325–334. [PubMed: 15690988]
8. Shen Y, Zelen M. Parametric estimation procedures for screening programmes: Stable and nonstable disease models for multimodality case finding. Biometrika. 1999; 86:503–515.
9. Johnson, NL.; Kotz, S.; Balakrishnan, N. Continuous Univariate Distributions. 2. Wiley; New York: 1995. p. 248
10. [Last access on December 28, 2012] http://en.wikipedia.org/wiki/Beta_prime_distribution
11. McDonald JB, Xub YJ. A generalization of the beta distribution with applications. Journal of Econometrics. 1995; 66:133–152. Corr: 1995;69:427–428.
12. Coelho CA, Mexia JT. On the distribution of the product and ratio of independent generalized gamma-ratio random variables. Sankhy : The Indian Journal of Statistics. 2007; 69(Part 2):221–255.
13. Pham-Gia T. Exact distribution of the generalized Wilks's statistic and applications. Journal of Multivariate Analysis. 2008; 99:1698–1716.10.1016/j.jmva.2008.01.021
14. Pham-Gia T, Turkkan N. Operations on the generalized- F variables and applications. Statistics. 2002; 36:195–209. CIsid: 236851.
15. Bekker A, Roux J, Pham-Gia T. The type I distribution of the ratio of independent “Weibullized” generalized beta-prime variables. Statistical Papers. 2009; 50:323–338.10.1007/s00362-007-0083-2
16. Pregibon D. Logistic regression diagnostics. Annals of Statistics. 1981; 9:705–24.
17. Adewale AJ, Xu X. Robust designs for generalized linear models with possible overdispersion and misspecified link functions. Computational Statistics and Data Analysis. 2010; 54:875–90.
18. Zelterman D, Chen C-F. Homogeneity tests against central mixture alternatives. Journal of the American Statistical Association. 1988; 83:179–182.
19. Zelterman D, Tulupyev AL, Suvorova AV, Paschenko AE, Musina VF, Tulupyeva TV, Krasnoselskikh TV, Grau L, Heimer R. Processing length bias of time intervals between the last episode and the interview in Gamma-Poisson models of behavior. SPIIRAS Proceedings. 2011; 16:160–185. (In Russian). Available at: <http://mi.mathnet.ru/eng/trspy/v16/p160>.

Appendix: Details of the Marginal Distribution

Lemma 1

For any observed values $\mathbf{t} = \{t_i > 0; i = 1, \dots, n\}$ there exist critical values $(\hat{\alpha}, \hat{\sigma})$ of the beta prime log-likelihood of (5) with $\hat{\alpha} > 0$ and $\hat{\sigma} > 0$.

Proof

The log-likelihood of (5) is

$$\begin{aligned}\Lambda(\alpha, \sigma) &= \sum \log f(t_i | \alpha, \sigma) \\ &= n \log \alpha + n \log(\alpha + 1) - 2n \log \sigma - (\alpha + 2) n S(\mathbf{t}/\sigma) + \sum_i \log(t_i)\end{aligned}\quad (10)$$

where

$$S = S(\mathbf{t}/\sigma) = n^{-1} \sum_i \log(1 + t_i/\sigma)$$

is positive valued.

The equation $\Lambda/\alpha = 0$ is quadratic in α . The smaller root is negative and the larger root, denoted by

$$\hat{\alpha} = 1/S - 1/2 + (2S)^{-1}(S^2 + 4)^{1/2}$$

satisfies $\hat{\alpha} > 1/S$ and so it must be positive valued.

Similarly, the equation $\Lambda/\sigma = 0$ requires a solution in $\hat{\sigma}$ of the equation

$$\sum_i^n \left(\frac{t_i}{\hat{\sigma} + t_i} \right) = 2n/(2 + \hat{\alpha}).$$

This summation is a continuous, monotone decreasing function of $\hat{\sigma}$ defined on $(0, +\infty)$ taking values from n down to zero. We showed that $\hat{\alpha} > 0$ so there must be a solution to this equation in finite $\hat{\sigma} > 0$.

The observed information for the likelihood (10) has entries

$$\begin{aligned}-\frac{\partial^2 \Lambda}{\partial \alpha^2} &= n/\alpha^2 + n/(\alpha + 1)^2 \\ -\frac{\partial^2 \Lambda}{\partial \sigma^2} &= -2n\sigma^{-2} + 2(\alpha + 2)\sigma^{-2} \sum \left(\frac{t_i}{\sigma + t_i} \right) - (\alpha + 2)\sigma^{-2} \sum \left(\frac{t_i}{\sigma + t_i} \right)^2 \\ -\frac{\partial^2 \Lambda}{\partial \alpha \partial \sigma} &= \sigma^{-1} \sum \left(\frac{t_i}{\sigma + t_i} \right)\end{aligned}$$

and the expected information is

$$\begin{aligned} -E \frac{\partial^2 \Lambda}{\partial \alpha^2} &= n/\alpha^2 + n/(\alpha+1)^2 \\ -E \frac{\partial^2 \Lambda}{\partial \sigma^2} &= 2n(\alpha - \sigma - 2\alpha\sigma + 2) / \{\sigma^2(\alpha+2)\} \\ -E \frac{\partial^2 \Lambda}{\partial \alpha \partial \sigma} &= -2n/(\alpha+2). \end{aligned}$$

Figure 1. Histogram of marginal times since last abuse of alcohol. The single extremely large observation appears as a star in Fig. 3.

Figure 2.

Beta prime density functions, with shape parameters α as specified and all with scale parameter $\sigma = 1$. These parameter values are in the range of those estimated in Section 4. Larger values of α are associated with more recent, and more frequent events.

Figure 3. Regression diagnostics for beta prime regression model (9). Observed last-event times of $t_i = 0$ (recoded as $t = .5$) and $t_i = 1$ are displayed throughout as open circles. The star indicates a single extremely large observation.

Figure 4.

Perturbation (as a percentage of relative change) of estimated parameter values in model (9) for different coding of values for $t = 0$ and $t = 1$. In this study, values for $t = 1$ were coded as $2 \times$ those of $t = 0$.

Table 1

Fitted regression coefficients, standard errors, and statistical significance for all marginal models and the joint regression model. The model for $a_i = \exp(\beta x_i)$ is given at (7) and the same estimated value of scale parameter σ is used for all subjects.

Covariate		Full model	Marginal models	Age and Coping
Sex	$\hat{\beta}$	-.081	.019	
	$SE(\hat{\beta})$	.132	.120	
	p	.5	.9	
Age	$\hat{\beta}$	-.011	-.015	-.012
	$SE(\hat{\beta})$	.008	.007	.008
	p	.17	.041	.102
Coping	$\hat{\beta}$	-.300	-.275	-.244
	$SE(\hat{\beta})$	.110	.103	.104
	p	.006	.008	.019
New	$\hat{\beta}$	.012	.028	
	$SE(\hat{\beta})$	.027	.024	
	p	.7	.2	
Defense/100	$\hat{\beta}$	-.025	-.018	
	$SE(\hat{\beta})$	.311	.305	
	p	.9	.9	
Risky/100	$\hat{\beta}$	.661	.413	
	$SE(\hat{\beta})$	.410	.389	
	p	.15	.3	
Intercept	$\hat{\beta}$	.220	—	.180
	$SE(\hat{\beta})$	.410	—	.294
Scale	$\hat{\sigma}$	1.167	—	1.150
	$SE(\hat{\sigma})$	.191	—	.189
Log-Likelihood-1100		53.06	—	54.65

Table 2

Fitted regression coefficients and log likelihood for models of $\sigma_i = \exp(\gamma'x_i)$ and fixed α for all subjects.

Covariate		Full model	Marginal models	Age, New, Coping, Risky
Sex	$\hat{\gamma}$	-.024	-.177	
	$SE(\hat{\gamma})$	.25	.233	
	p	.9	.4	
Age	$\hat{\gamma}$	-.024	.029	.024
	$SE(\hat{\gamma})$	.015	.015	.014
	p	.09	.048	.096
Coping	$\hat{\gamma}$	.548	.541	.544
	$SE(\hat{\gamma})$	.202	.199	.196
	p	.007	.007	.006
New	$\hat{\gamma}$	-.094	-.135	-.093
	$SE(\hat{\gamma})$	.051	.047	.051
	p	.067	.004	.068
Defense/100	$\hat{\gamma}$	.128	-.339	
	$SE(\hat{\gamma})$	.643	.642	
	p	.8	.6	
Risky/100	$\hat{\gamma}$	-1.290	-1.880	-1.304
	$SE(\hat{\gamma})$	.894	.796	.843
	p	.15	.018	.12
Shape	$\hat{\alpha}$	.540	—	.540
	$SE(\hat{\alpha})$	.046	—	.046
Intercept	$\hat{\gamma}_0$	-.873	—	-.845
	$SE(\hat{\gamma}_0)$	.746	—	.719
Log-Likelihood-1100		49.21	—	49.24

Table 3

Fitted models for $\alpha_i = \exp(\beta'x_i)$ and $\sigma_i = \exp(\gamma'x_i)$. These models use the explanatory variables Coping in α and Age, Coping, and/or New individually and jointly in σ . Fitted parameters for model (9) are in the indicated column.

Model (9)											
α_i	Intercept	$\hat{\beta}_0$	-.179	-.379	-.138	-.422	-.152	-.336	-.371		
		SE($\hat{\beta}_0$)	.208	.270	.204	.271	.205	.268	.269		
Coping		$\hat{\beta}$	-.258	-.153	-.275	-.120	-.262	-.164	-.139		
		SE($\hat{\beta}$)	.104	.142	.101	.143	.102	.141	.142		
		p	.013	.3	.006	.4	.010	.2	.3		
σ_i	Intercept	$\hat{\gamma}_0$	-.623	-.464	1.168	-1.323	.504	.621	-.145		
		SE($\hat{\gamma}_0$)	.468	.502	.378	.689	.604	.616	.799		
Age		$\hat{\gamma}$	.025			.027	.021		.022		
		SE($\hat{\gamma}$)	.015			.014	.014		.014		
		p	.084			.065	.15		.12		
Coping		$\hat{\gamma}$	.337			.375		.311	.340		
		SE($\hat{\gamma}$)	.272			.270		.276	.237		
		p	.2			.16		.3	.2		
New		$\hat{\gamma}$			-.136		-.129	-.135	-.127		
		SE($\hat{\gamma}$)			.046		.046	.046	.046	.046	
		p			.003		.005	.004	.006		
Log likelihood-1100			54.63	55.29	51.74	53.65	50.74	51.10	49.96		