Identifiability of the misspecified split hazard models

Sanjiv Jaggia

Unlike standard models, a split population hazard model allows the exit probability to be less than one. Although conceptually attractive, split models are prone to identification problems. In the reduced form estimation of the hazard function, the influence of split may not be distinguishable from that of neglected heterogeneity. For illustration, I use Monte Carlo simulations to highlight the problem of interpreting the structural parameters of the split Weibull and the Weibull-gamma models.

I. Introduction

Hazard rate models are used to study the instantaneous probability (hazard) of a transition from one state to another, given that the transition has not already occurred. An implicit assumption in standard models is that of certain exit implying that all observations will eventually experience the event of interest if the observation period is sufficiently prolonged. In an application of unemployment durations, it implies that all unemployed individuals will eventually find employment. In studying criminal recidivism, Schmidt and Witte (1989) argue that some criminals are 'cured' in that they will never commit another crime that sends them back to prison. They introduce a split hazard model that takes into account the possibility that the transition from one state to another may *never* occur.[1]

In this article, I show that although split models have an obvious intuitive appeal in social sciences, they are prone to certain identification problems. In particular, a split parameter can spuriously be influenced by the misspecification of the functional form of the underlying hazard function. Similarly, an incorrect functional form of the hazard may be inferred when the hazard model is split. For illustration, I show that when the underlying model is Weibull-gamma, the estimated split Weibull model spuriously indicates that a fraction of observations will never experience an exit. Similarly, researchers may confuse split data with the presence of neglected heterogeneity in the model. The allowance for both split and neglected heterogeneity may just be compensating for a restrictive Weibull specification that only allows monotonic hazards. It is difficult to discriminate between the split Weibull and the Weibull-gamma models since the reduced form of both models permit an 'inverted U' shape of the hazard. This result is highlighted with Monte Carlo experiments. I argue that although the reduced forms are somewhat similar, the interpretation of the results for the two models can be quite different.

[1] In biostatistics, these models, referred to as cure models, allow for a cured fraction of individuals who will never experience a reoccurrence of disease. For applications in economics and finance, see Bandopadhyaya and Jaggia (2001), DeYoung (2003), Mavromaras and Orme (2004), Chang and Yeh (2007), Madden (2007) etc.

II. Background

Hazard models are represented in terms of the density, survivor and hazard functions, denoted by $f(t; X)$, $S(t; X)$ and $h(t; X)$ respectively.[2] These functions, conditional on the vector of characteristics X, are defined as follows:

$$h(t; X) = \frac{f(t; X)}{S(t; X)}; \quad S(t; X) = P(T > t);$$
$$f(t; X) = -\frac{\partial S(t; X)}{\partial t} \tag{1}$$

Let C be an indicator variable that equals 1 if the duration is complete and 0 if it is right censored. The following log-likelihood function uses the logs of $f(t; X)$ for completed and $S(t; X)$ for censored observations:

$$\ln L = \sum_{i=1}^{N} C_i(\ln f(t_i; X_i)) + (1 - C_i)\ln(S(t_i; X_i)) \tag{2}$$

An implicit assumption made in the above formulation is that of certain exit, implying that $S(t; X) \to 0$ as $t \to \infty$. In other words, an event (transition from one state to another) will take place if the observation period is sufficiently long. A split hazard model takes into account the possibility that for some observations, the exit may not happen.[3] The resulting model estimates the hazard parameters along with the split parameter, δ, which allows the probability of eventual exit to be less than one.

As mentioned earlier, we typically observe completed as well as censored observations. For some censored observations, the exit may never happen. Let u be a latent variable that equals 1 for an eventual exit and 0 otherwise. Further let $P(u = 1) = \delta$, where $\delta \leq 1$ represents the probability of eventual exit. If the event occurs, we have $C = 1$ and $u = 1$. The appropriate contribution for such an observation is

$$P(u = 1)f(t; X, u = 1) = \delta f(t; X, u = 1) \tag{3}$$

For a censored observation, we entertain two possibilities: (a) the event would occur if the observation period is prolonged and (b) the event would never occur. Specifically, the contribution of a censored observation is

$$P(u = 0) + P(u = 1)S(t; X, u = 1)$$
$$= 1 - \delta + \delta S(t; X, u = 1) \tag{4}$$

The corresponding log-likelihood function of a split hazard model is

$$\ln L = \sum_{i=1}^{N} C_i(\ln \delta + \ln f(t_i; X_i)) + (1 - C_i)$$
$$\times \ln(1 - \delta + \delta S(t_i; X_i)) \tag{5}$$

where $f(t; X) = f(t; X, u = 1)$ and $S(t; X) = S(t; X, u = 1)$.

Split models are known to be prone to certain inherent identification problems. First, a censored observation suggests that the event of interest has not yet happened, but it is not clear if the event would have occurred had the observation period been prolonged. Second, researchers often formulate δ as a logistic function of the X variables, which are also used in the hazard. Since economic theory usually provides no direction, it is difficult to identify the influence of the same variable on both the hazard and the eventual exit probability.

Another issue that has not been discussed in the literature is that the split parameter can spuriously compensate for a misspecified functional form of the hazard function. Similarly, split data may be observationally equivalent to a heterogeneous model. This issue becomes obvious once we derive the reduced form of the split hazard as follows:

$$h(t; X) = \frac{\delta f(t; X, u = 1)}{1 - \delta + \delta S(t; X, u = 1)} \tag{6}$$

As we can see from Equation (6), with δ less than 1, the split model forces the hazard to decline, especially at high t-values. This condition will hold irrespective of the functional form of the underlying hazard. In the following section, we will show how the above property causes the identification problem between the reduced forms of the split Weibull and the Weibull-gamma models.

III. The Weibull Model

For parametric estimation, we must specify *a priori* the functional form of the hazard function. The hazard of a commonly used Weibull distribution is

$$h(t; X) = \mu\alpha t^{\alpha-1} \tag{7}$$

The scale parameter $\mu = \exp(X\beta)$, whereas the shape parameter α allows monotonically increasing $\alpha > 1$ and decreasing $\alpha < 1$ hazard functions.

[2] See Kiefer (1988) and Lancaster (1990).
[3] See Schmidt and Witte (1989) and Bandopadhyaya and Jaggia (2001) for details.

If the Weibull specification is deemed inappropriate, an appropriate formulation is $\mu = v\exp(X\beta)$ where v accounts for neglected heterogeneity. As v is not observable, the unconditional survivor function is

$$S(t; X) = \int_0^\infty \exp(-v\exp(X\beta)t^\alpha) f(v)\mathrm{d}v \qquad (8)$$

Using a convenient gamma mixing distribution, with a unit mean and variance equal to σ^2, we can easily derive

$$S(t; X) = \left[1 + \sigma^2 \mu t^\alpha\right]^{-1/\sigma^2} \qquad (9)$$

The corresponding reduced form hazard for a Weibull-gamma model is derived as

$$h(t; X) = \frac{\mu\alpha t^{\alpha-1}}{1 + \sigma^2 \mu t^\alpha} \qquad (10)$$

Note that the above function allows an 'inverted-U' shape if $\sigma > 1$ and $\sigma^2 > 1$.[4] On the other hand, the reduced form of the split Weibull model (see Equation 6) is

$$h(t; X) = \frac{\delta\alpha\mu t^{\alpha-1}\exp(-\mu t^\alpha)}{1 - \delta + \delta\exp(-\mu t^\alpha)} \qquad (11)$$

Although the Weibull formulation only allows monotonic hazards, with the split parameter $\delta < 1$, it can also accommodate an 'inverted U' shape if $\alpha > 1$. While different in the structural formulations, the reduced forms of the split Weibull as well as the Weibull gamma hazards are somewhat similar.

IV. Monte Carlo Experiments

Monte Carlo experiments are conducted to highlight the identification problem discussed above. For all experiments, a sample of 200 observations, repeated 2000 times, is used to compare the performance of three models, namely, the estimated Weibull, Weibull-gamma and split Weibull models. These models are compared on the basis of their parameter estimates as well as the estimated hazards evaluated at various points in time. True models used in simulations are the (a) Weibull-gamma, (b) split Weibull and (c) split Weibull-gamma models. In each case, the base model is a Weibull-gamma with $\mu = v\exp(-5 + X_1 - 0.5X_2)$ and the duration dependence parameter, $\alpha = 2$. The variables X_1 and X_2 are drawn from a standard normal distribution and are held fixed for all experiments. In order to incorporate neglected heterogeneity, v is drawn from the gamma distribution with a unit mean, $E(v) = 1$, and variance, $\sigma_v^2 = 1$ for small and $\sigma_v^2 = 2$ for large heterogeneity.[5] In order to introduce split, I randomly select 20% of the observations and make their durations infinite. In other words, I use the split parameter, $\delta = 0.80$, implying that only 80% of the observations will eventually conclude. Finally, in order to incorporate censoring, I impose thresholds on data observation periods so that about 38–42% of the observations are right censored. Remember that with split data, a censored observation denotes that an event will either occur beyond the censoring point or that it will never occur.

Weibull-gamma model

Here, I simulate data according to the Weibull-gamma specification with small ($\sigma_v^2 = 1$) and large ($\sigma_v^2 = 2$) levels of neglected heterogeneity. As mentioned earlier, simulated data are used to estimate the Weibull, Weibull-gamma and split Weibull models. The means of the estimated parameters and their respective t-statistics are presented in Tables 1 and 3. The corresponding hazard at various points in time, evaluated at the mean estimates, is presented in Tables 2 and 4, respectively. As expected, since the Weibull model does not accommodate a nonmonotonic hazard, it spuriously underestimates the duration dependence parameter α. Given the true $\alpha = 2$, the mean estimates are $\hat{\alpha} = 1.51$ for small and $\hat{\alpha} = 1.24$ for large heterogeneity. The correctly specified Weibull-gamma model estimates the parameters precisely and the corresponding hazard captures the true shape. The most interesting result pertains to the significance of the split parameter of the split Weibull model. The mean estimate of the split parameter, $\hat{\delta} = 0.88\,(0.83)$ for small (large) heterogeneity, spuriously implies that only 88% (83%) of the observations will eventually exit while the remaining 12% (17%) will never exit.

As noted above, the Weibull specification is restrictive since it does not allow a nonmonotonic hazard. However, the reduced forms of the Weibull-gamma as well as the split Weibull models permit an 'inverted U' shape (Equations 10 and 11). Tables 2 and 4 highlight the difficulty of

[4] Jaggia and Thosar (1995) suggest that the mixing distribution is used not only to compensate for omitted factors, but also to correct for an overly restrictive Weibull hazard function.

[5] Note that for $\sigma_v^2 = 1$, the Weibull-gamma model specializes to a log-logistic model. Further, for $\sigma_v^2 = 0$, the model reduces to a basic Weibull with no heterogeneity.

Table 1. Parameter estimates (true model: Weibull-gamma with small variance)

Parameters	True values	Weibull	Weibull-gamma	Split Weibull
Constant	−5.00	−4.24 (−13.24)	−5.34 (−9.61)	−4.54 (−12.63)
X_1	1.00	0.69 (7.07)	1.13 (5.18)	0.82 (7.59)
X_2	−0.50	−0.37 (−3.97)	−0.65 (−3.58)	−0.46 (−4.34)
α	2	1.51 (4.28)	2.19 (3.88)	1.74 (4.92)
σ^2	1	NA	1.33 (2.30)	NA
δ	1	NA	NA	0.88 (−2.54)

Notes: The table contains the means of the estimated parameters and the corresponding t-values, in parentheses, from 2000 simulations. α represents the shape parameter of the Weibull model, σ^2 is the variance of the heterogeneity term and δ is the split parameter. For α and δ, the t-statistics are evaluated at 1.

Table 2. Estimated hazards (true model: Weibull-gamma with small variance)

Duration	True hazard	Weibull	Weibull-gamma	Split Weibull
5	0.054	0.047	0.054	0.049
10	0.077	0.066	0.078	0.076
15	0.078	0.081	0.076	0.089
20	0.071	0.094	0.066	0.084
25	0.064	0.105	0.058	0.061
30	0.057	0.116	0.050	0.033
35	0.051	0.125	0.044	0.013
40	0.045	0.134	0.039	0.004

Table 3. Parameter estimates (true model: Weibull-gamma with large variance)

Parameters	True values	Weibull	Weibull-gamma	Split Weibull
Constant	−5.00	−3.73 (−13.29)	−5.21 (−9.27)	−4.03 (−12.73)
X_1	1.00	0.51 (5.34)	1.15 (4.68)	0.71 (6.49)
X_2	−0.50	−0.23 (−2.58)	−0.55 (−2.89)	−0.32 (−3.12)
α	2	1.24 (2.55)	2.21 (3.61)	1.51 (4.04)
σ^2	2	NA	2.21 (2.94)	NA
δ	1	NA	NA	0.83 (−3.51)

Notes: The table contains the means of the estimated parameters and the corresponding t-values, in parentheses, from 2000 simulations. α represents the shape parameter of the Weibull model, σ^2 is the variance of the heterogeneity term and δ is the split parameter. For α and δ, the t-statistics are evaluated at 1.

Table 4. Estimated hazards (true model: Weibull-gamma with large variance)

Duration	True hazard	Weibull	Weibull-gamma	Split Weibull
5	0.047	0.042	0.056	0.046
10	0.055	0.050	0.064	0.060
15	0.049	0.055	0.054	0.064
20	0.042	0.059	0.045	0.059
25	0.035	0.063	0.037	0.046
30	0.031	0.066	0.032	0.031
35	0.027	0.068	0.028	0.018
40	0.024	0.070	0.024	0.009

Table 5. Parameter estimates (true model: split Weibull with no heterogeneity)

Parameters	True values	Weibull	Weibull-gamma	Split Weibull
Constant	−5.00	−4.18 (−13.35)	−5.95 (−9.66)	−5.10 (−12.59)
X_1	1.00	0.58 (6.14)	1.28 (5.47)	1.02 (8.50)
X_2	−0.50	−0.31 (−3.32)	−0.65 (−3.54)	−0.51 (−4.68)
α	2	1.47 (4.15)	2.52 (4.55)	2.04 (6.24)
σ^2	0	NA	1.89 (3.14)	NA
δ	0.80	NA	NA	0.80 (−4.53)

Notes: The table contains the means of the estimated parameters and the corresponding t-values, in parentheses, from 2000 simulations. α represents the shape parameter of the Weibull model, σ^2 is the variance of the heterogeneity term and δ is the split parameter. For α and δ, the t-statistics are evaluated at 1.

Table 6. Estimated hazards (true model: split Weibull with no heterogeneity)

Duration	True hazard	Weibull	Weibull-gamma	Split Weibull
5	0.048	0.046	0.055	0.048
10	0.085	0.063	0.079	0.087
15	0.093	0.077	0.071	0.095
20	0.062	0.088	0.060	0.062
25	0.024	0.098	0.050	0.022
30	0.005	0.106	0.043	0.005
35	0.001	0.114	0.037	0.001
40	0.000	0.122	0.033	0.000

discriminating between these two models since the Weibull model clearly does not capture the essential feature of the data. The estimated hazard at various points in time of the split Weibull model captures the basic shape of the true Weibull-gamma process although its decline is steeper at high durations. The two models, however, have very different interpretation of the structural parameters. For instance, although all observations in the sample do eventually exit, the split model suggests that a portion of the observations will never do so.

Split Weibull model

The results of the three estimated models when the true model is the split Weibull are presented in Tables 5 and 6. Again, the Weibull model underestimates the duration dependence parameter, with the average estimate of the duration dependence parameter, $\hat{\alpha} = 1.47$. As expected, the parameter estimates of the estimated the split Weibull model are consistent with their true parameter values. The Weibull-gamma model, on the other hand, suggests neglected heterogeneity when none existed. For instance, the true variance, $\sigma^2 = 0$, is spuriously estimated as $\hat{\sigma}^2 = 1.89$. The average hazard of the Weibull-gamma accommodates the inverted 'U' shape of the hazard function; however it does not drop as steeply as the true hazard. As before, the Weibull model is unable to capture the nonmonotonic hazard function.

Split Weibull-gamma model

Here the same three estimated models are compared using simulated data from the split Weibull-gamma model, with $\delta = 0.8$ with $\sigma_v^2 = 1$ for small and $\sigma_v^2 = 2$ for large heterogeneity; see Tables 7–10 for results. Note that none of the three estimated models are appropriate. The Weibull model is clearly inappropriate. The Weibull-gamma model, that ignores split, exaggerates the presence of neglected heterogeneity. The average variances of the heterogeneity term are estimated as $\hat{\sigma}_v^2 = 3.21$ and $\hat{\sigma}_v^2 = 4.10$ when the true variances equals 1 and 2, respectively. Similarly the split Weibull model, that ignores neglected heterogeneity, results in the average estimate of the split parameter as $\hat{\delta} = 0.74$ ($\hat{\delta} = 0.71$) with small (large) heterogeneity when the true value is $\delta = 0.8$ Although the Weibull-gamma and split Weibull models are both misspecified, they are able to somewhat capture the true shape of the hazard; the Weibull gamma model is preferable with large neglected heterogeneity.

Table 7. Parameter estimates (true model: split Weibull-gamma with small variance)

Parameters	True values	Weibull	Weibull-gamma	Split Weibull
Constant	−5.00	−3.70 (−13.10)	−5.67 (−8.85)	−4.19 (−12.35)
X_1	1.00	0.42 (4.48)	1.23 (4.49)	0.74 (6.47)
X_2	−0.50	−0.17 (−1.91)	−0.51 (−2.31)	−0.27 (−2.48)
α	2	1.15 (1.67)	2.41 (3.70)	1.57 (4.37)
σ^2	1	NA	3.21 (3.34)	NA
δ	0.80	NA	NA	0.74 (−5.65)

Notes: The table contains the means of the estimated parameters and the corresponding t-values, in parentheses, from 2000 simulations. α represents the shape parameter of the Weibull model, σ^2 is the variance of the heterogeneity term and δ is the split parameter. For α and δ, the t-statistics are evaluated at 1.

Table 8. Estimated hazards (true model: split Weibull-gamma with small variance)

Duration	True hazard	Weibull	Weibull-gamma	Split Weibull
5	0.037	0.035	0.049	0.040
10	0.040	0.039	0.054	0.052
15	0.033	0.042	0.044	0.054
20	0.026	0.044	0.035	0.045
25	0.020	0.045	0.029	0.032
30	0.016	0.046	0.024	0.019
35	0.013	0.048	0.021	0.010
40	0.011	0.049	0.018	0.004

Table 9. Parameter estimates (true model: split Weibull-gamma with large variance)

Parameters	True values	Weibull	Weibull-gamma	Split Weibull
Constant	−5.00	−3.55 (−13.01)	−5.68 (−8.33)	−3.90 (−12.20)
X_1	1.00	0.41 (4.41)	1.29 (4.25)	0.68 (5.97)
X_2	−0.50	−0.12 (−1.26)	−0.45 (−1.85)	−0.18 (−1.58)
α	2	0.98 (−0.26)	2.30 (3.16)	1.33 (2.89)
σ^2	2	NA	4.10 (3.26)	NA
δ	0.80	NA	NA	0.71 (−5.88)

Notes: The table contains the means of the estimated parameters and the corresponding t-values, in parentheses, from 2000 simulations. α represents the shape parameter of the Weibull model, σ^2 is the variance of the heterogeneity term and δ is the split parameter. For α and δ, the t-statistics are evaluated at 1.

Table 10. Estimated hazards (true model: split Weibull-gamma with large variance)

Duration	True hazard	Weibull	Weibull-gamma	Split Weibull
5	0.037	0.027	0.039	0.030
10	0.040	0.026	0.040	0.034
15	0.033	0.026	0.032	0.035
20	0.026	0.026	0.026	0.032
25	0.020	0.026	0.021	0.028
30	0.016	0.026	0.018	0.023
35	0.013	0.026	0.016	0.018
40	0.011	0.026	0.014	0.013

V. Conclusion

Hazard models typically assume that every agent in the population is susceptible to the event in study and will eventually experience the event if the observation period is sufficiently long. This assumption is appropriate in medical and engineering sciences since eventually all machines break down and all patients die. In economics, however, not everyone finds a job and not all firms file for bankruptcy. The implausibility of this assumption has prompted some researchers to use split models that allow some agents not to be susceptible and, therefore, will never experience the event of interest. Although split models have the obvious intuitive appeal in applications in social sciences, they are prone to identification problems. In particular, in the reduced form estimation of the hazard function, the influence of split may not be distinguishable from that of neglected heterogeneity. Monte Carlo simulations suggest that it is difficult to discriminate between the split Weibull and the Weibull-gamma models since the reduced forms of these models are somewhat similar.

It is not unreasonable to expect both neglected heterogeneity and split in the hazard model applications in social sciences. Since the effect of the two issues can be confounded, as shown in the article, care must be exercised in interpreting the models. If correct interpretation of the structural parameters is essential, it is important that a generally specified hazard model be used. For instance, in order to correctly capture the split parameter, the hazard function needs to be correctly specified with allowance made for neglected heterogeneity. Given the uncertainty regarding the true hazard, it is perhaps better to analyse the heterogeneous and split models in their reduced form formulations since they may also have been compensating for an overly restrictive hazard function.

References

Bandopadhyaya, A. and Jaggia, S. (2001) An analysis of second time around bankruptcies using split population duration models, *Journal of Empirical Finance*, **8**, 201–18.

Chang, H.-L. and Yeh, T.-H. (2007) Exploratory analysis of motorcycle holding time heterogeneity using a split-population duration model, *Transportation Research Part A: Policy and Practice*, **41**, 587–96.

DeYoung, R. (2003) The failure of new entrants in commercial banking markets: a split population duration analysis, *Review of Financial Economics*, **12**, 7–33.

Jaggia, S. and Thosar, S. (1995) Contested tender offers: an estimate of the hazard function, *Journal of Business and Economic Statistics*, **13**, 113–19.

Kiefer, N. M. (1988) Econometric duration data and hazard functions, *Journal of Economic Literature*, **25**, 646–79.

Lancaster, T. (1990) *The Econometric Analysis of Transition Data*, Cambridge University Press, New York.

Madden, D. (2007) Tobacco taxes and starting and quitting smoking: does the effect differ by education?, *Applied Economics*, **39**, 613–27.

Mavromaras, K. G. and Orme, C. D. (2004) Temporary layoffs and split population models, *Journal of Applied Econometrics*, **19**, 49–67.

Schmidt, P. and Witte, A. D. (1989) Predicting criminal recidivism using split population survival time model, *Journal of Econometrics*, **40**, 141–59.