Paper 195-30
Single Event Timing Model with Heterogeneity: Predicting the Time of the First Home Purchase by Immigrants

Eugene Brusilovskiy ● The Wharton School, University of Pennsylvania
Philadelphia, PA

Introduction

It is well known that unobserved heterogeneity can produce misleading estimates in survival analysis (see for example, Allison, 1995, p.233). The goal of this paper is to show how to estimate and apply a single event timing model with heterogeneity to real data with SAS®. We consider three different models:
- An Exponential-gamma model, when individual-level time-to-event is exponentially distributed, and the event-rate is gamma-distributed
- A Weibull-gamma model, when individual-level time-to-event has a Weibull distribution, and the event-rate has a gamma-distribution (Exponential-gamma is a particular case of this model)
- A latent class Weibull model that allows for heterogeneity in both shape and scale parameters. Thus, the assumption is that several segments of individuals exist with their own fixed but unknown values of shape and scale parameters.

All unknown parameters in the models with heterogeneity were estimated by the maximum likelihood method, using the nonlinear optimization procedure *PROC NLP*. For illustration, real data from the American Housing Survey were used to develop and compare survival analysis type models to analyze the time of the first house purchase by immigrants. Exploratory data analysis and visualization were implemented with *PROC LIFETEST* (testing for differences in survivor functions and survivor function plots) and with *PROC RELIABILITY* (probability plots). Competitors for three models with heterogeneity were accelerated failure time models that were produced by *PROC LIFEREG*, and proportional hazard models that were produced by *PROC PHREG*.

Data

The data are taken from the American Housing Survey (AHS) that is available at www.huduser.org. Conducted by the U.S. Bureau of the Census, AHS is the largest regular national housing sample survey in the United States. In this paper, we use the national level data from 2003 survey, limiting the sample to 989 foreign-born naturalized citizens who either rent an apartment or own a house. Each individual/householder came to the US within the last 20 years, and if he moved into the current house, we assume that this is the *first* house that he has purchased in the US. So, the variable of interest here purchasing time of the first house in the USA (measured in months), and explanatory variables were sex (Male, Female), Spanish origin of householder (Yes, No), Status of Unit (Own, Rent), Race (White, Asian, Other), Marital Status (Married, Not-married), and the householder's Age in years. The variable Status of Unit helped us to form data in the form (t_i, δ_i), i =1,2,…,n, where δ_i is an indicator variable of house purchasing: δ_i =1 means that event time was observed for an individual i,, (a house was purchased), and δ_i = 0 means that event time was not observed for that individual. Thus, we had right-censored data.

Variables like income, socio-economic status of people in the neighborhood, house cost, number of bedrooms, and other relevant information were not available. The impact of unobserved heterogeneity (the bias caused by not being able to include particular important explanatory variables) is much stronger in hazard models than in other types of regression models. The assumption of traditional hazard models - if

two individuals have identical values on the covariates, they also have identical hazard functions, is highly questionable for this data. Therefore, the problem of unobserved heterogeneity has to be addressed in the analysis of data under consideration.

Exploratory Data Analysis

The following tables give us a sense of the data structure used:

Table 1. Frequency Distribution of the Variable RACE

Race	Frequency	Percent	Cumulative Frequency	Cumulative Percent
ASIAN	306	30.94	306	30.94
OTHER	118	11.93	424	42.87
WHITE	565	57.13	989	100.00

Table 2. Frequency Distribution of the Variable Marital Status

Marital Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
NON_MARRIED	343	34.68	343	34.68
MARRIED	646	65.32	989	100.00

Table 3. Frequency Distribution of the Variable SEX

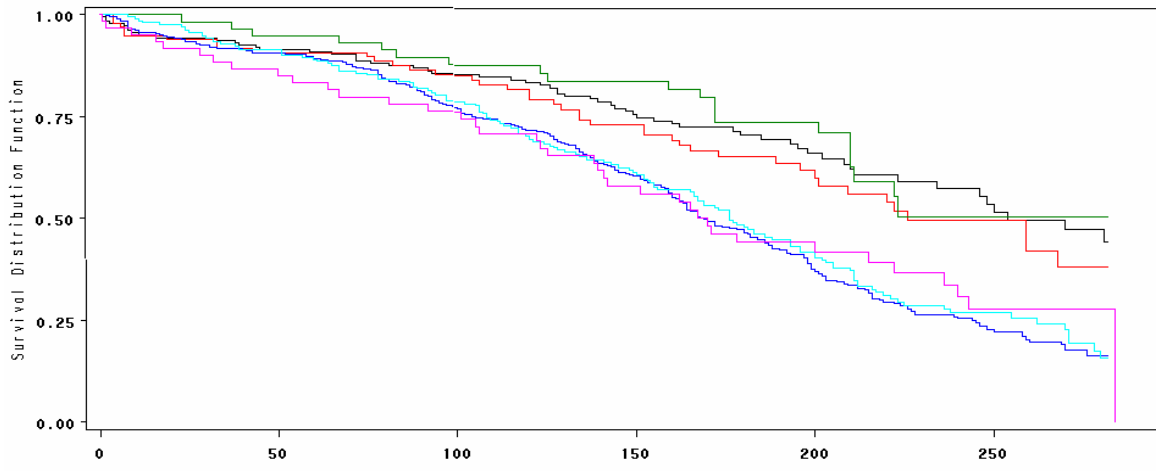
Sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
MALE	607	61.38	607	61.38
FEMALE	382	38.62	989	100.00

Table 4. Frequency Distribution of the Variable Spanish Origin

Spanish Origin	Frequency	Percent	Cumulative Frequency	Cumulative Percent
YES	352	35.59	352	35.59
NO	637	64.41	989	100.00

Survival functions were generated by *PROC LIFETEST* (SAS/STAT® software).

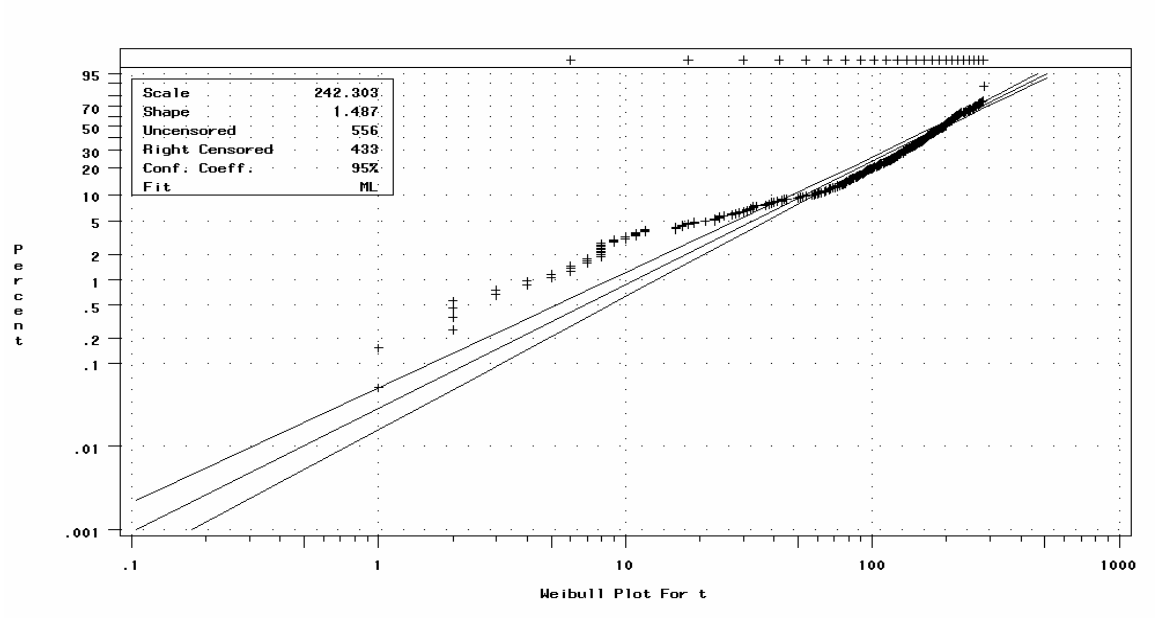
Graph 1. Survival Functions



Survival Function, Strata = Race and Marital Status, the upper branch of three curves corresponds to Marital Status=Non-Married, and the lower branch corresponds to Marital Status=Married. The hypothesis of equality over strata was rejected with $p < 0.001$.

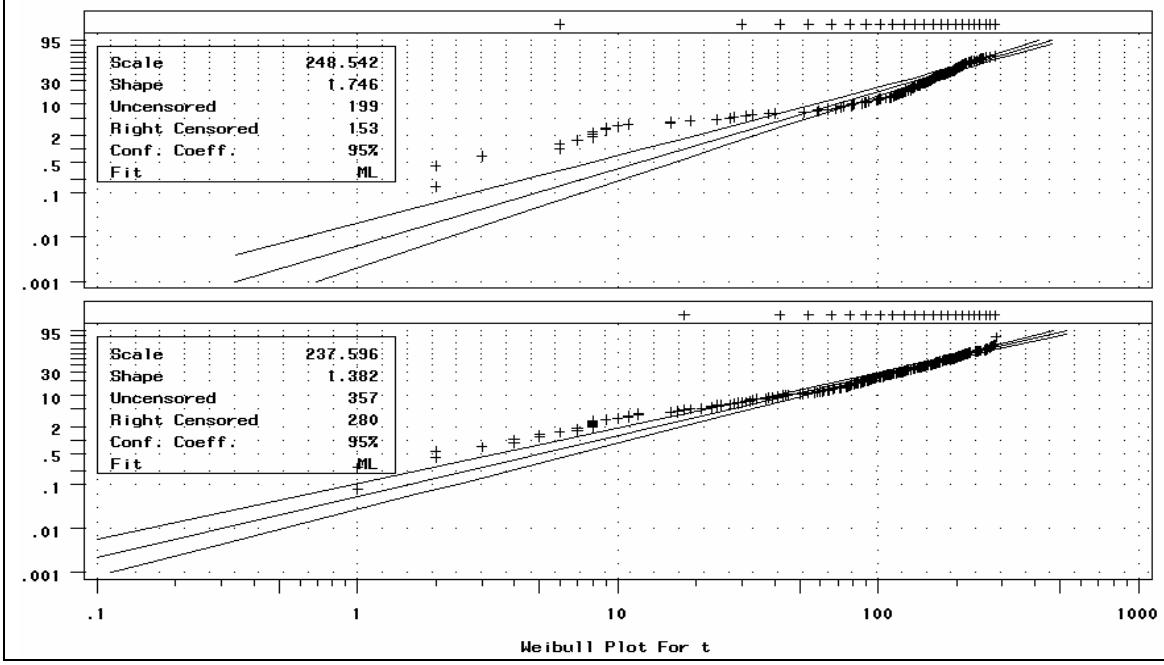
None of the distributions, provided by *PROC RELIABILITY* of SAS/QC® software (Weibull, exponential, extreme value, normal, log-normal, and log-logistic) fit the whole data. As an example, see Graph 2 – the probability plot for Weibull distribution.

Graph 2. Probability plot for Weibull distribution of time for first home purchase



Graph 3 below shows the probability plot for the Weibull distribution of time for the first home purchase for different values of the Spanish Origin variable.

Graph 3. The graph on top is for Spanish Origin = Yes, and the lower graph – for Spanish Origin = No



The segmentation of the original data by the variable Spanish Origin significantly improved the fit, but it is still poor. On the other hand, this exercise supports the idea that it makes sense to consider latent class Weibull models with heterogeneity – probably, a segmentation could help in the development of better models.

Accelerated Failure Time models

Accelerated failure time models were produced by *PROC LIFEREG* (SAS/STAT® software) for five different distributions.

Table 5. Log likelihoods for the accelerated failure time models with explanatory variables Race, Marital Status, Sex, Spanish Origin, and Age

Name of Distribution	Log Likelihood
Lognormal	-1154.682982
LLogistic	-1092.580208
Gamma (Best Fit)	-1036.719055
Exponential	-1107.721129
Weibull	-1056.834806

Since the log likelihoods are negative, lower (absolute) values suggest better fits. Thus, the best parametric regression model was with the Gamma distribution. Among all of the used covariates (explanatory variables), only Marital Status proved to be significant ($p<0.001$).

Proportional Hazard Models

Proportional hazard models were produced by *PROC HPREG* (SAS/STAT® software).
The log-likelihood for the proportional hazard regression model was 3426.98 using the same explanatory variables as in the accelerated failure time models. In this case, two explanatory variables were highly significant: Marital Status and Spanish Origin.

Models with Heterogeneity

Here, we introduce the Exponential-Gamma and Weibull-Gamma Models.

The Exponentia-Gamma Model Assumptions:
1. Individual level behavior of a homebuyer is characterized by the exponential distribution with rate parameter λ.
2. The values of λ are distributed across the population of homebuyers according to a gamma distribution with shape r and scale α parameters respectively.

The Weibul-Gamma Model Assumptions:
1. Individual level behavior of a homebuyer is characterized by the Weibull distribution with rate parameter λ and shape parameter c.
2. The values of λ are distributed across the population of homebuyers according to a gamma distribution with shape r and scale α parameters respectively.

The PDF and the CDF for the Wiebull-Gamma model are

$$f(t) = \frac{rct^{c-1}}{\alpha}\left(\frac{\alpha}{\alpha + t^c}\right)^{r+1}, \qquad F(t) = 1 - \left(\frac{\alpha}{\alpha + t^c}\right)^{r},$$

respectively (Bruce G.S. Hardie and Peter S. Fader, 2001). When c=1 we are getting corresponding functions for Exponential-Gamma model.

We denote t_i as the time at which the i^{th} homebuyer purchases the first house. However, we don't observe completed spells for all the individuals. Some of them never buy a house (right-censored). The likelihood function here is

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i}[1 - F(t_i)]^{1-\delta_i}$$

where $f(t_i)$ and $F(t_i)$ are PDF and CDF respectively. Therefore, after elementary transformations, the log-likelihood function for the Weibull-Gamma model is

$$LogLikelihood = \sum_i \left\{ \delta_i \log \frac{rct_i^{c-1}}{\alpha} + (\delta_i + r)\log\left(\frac{\alpha}{\alpha + t_i^c}\right)\right\},$$

and, in particular, if c=1 we will get the log likelihood function for Exponential-Gamma model.

The SAS code to get maximum likelihood estimates of Weibull-Gamma model is based on the usage of nonlinear programming procedure (*PROC NLP*) of the SAS/OR® software with default Newton-Raphson optimization algorithm:

Code Example 1. SAS Code for maximum likelihood estimates of the Weibull-Gamma Model

```
proc nlp data=EEE.DATA_EEE_MOD maxfunc=50000 maxiter=50000;
      max loglik;
            parms alpha=50, r=10,  c=2.5;
            bounds r > 2, alpha > 1e-12, c> 1;
      loglik = delta*log((r*c*t**(c-1))/alpha) +
              (delta + r)*log(alpha/(alpha + t**c));
run;
```

The parameter estimates were stable and almost independent of initial values, unlike in the situation when the objective function is not differentiable (Brusilovskiy, 2004).

Below, we look at the Latent Class Weibull Model. The assumptions for this model are as follows:

1. Individual level behavior of a homebuyer is characterized by the Weibull distribution with rate parameter λ and shape parameter c.
2. Heterogeneity is allowed in both λ and c.
3. Several segments of homebuyers exist, each with their own values for both of these parameters.

The CDF for two and three segments latent class Weibull models are

$$F_2(t) = \pi_1\left(1 - e^{-\lambda_1 t^{c_1}}\right) + (1 - \pi_1)\left(1 - e^{-\lambda_2 t^{c_2}}\right),$$

$$F_3(t) = \pi_1\left(1 - e^{-\lambda_1 t^{c_1}}\right) + \pi_2\left(1 - e^{-\lambda_2 t^{c_2}}\right) + (1 - \pi_1 - \pi_2)\left(1 - e^{-\lambda_3 t^{c_3}}\right)$$

(Bruce G.S. Hardie and Peter S. Fader(2001)). From here, probability density functions are

$$f_2(t) = F_3'(t) = \pi_1\lambda_1 c_1 t^{(c_1-1)} e^{-\lambda_1 t^{c_1}} + (1 - \pi_1)\lambda_2 c_2 t^{(c_2-1)} e^{-\lambda_2 t^{c_2}},$$

$$f_3(t) = F_3'(t) = \pi_1\lambda_1 c_1 t^{(c_1-1)} e^{-\lambda_1 t^{c_1}} + \pi_2\lambda_2 c_2 t^{(c_2-1)} e^{-\lambda_2 t^{c_2}} +$$

$$(1-\pi_1-\pi_2)\lambda_3 c_3 t^{(c_3-1)} e^{-\lambda_3 t^{c_3}}.$$

The log likelihood functions for two segments Weibull model is

$$LL_2 = \sum_{i=1}^{n}\left[\pi_1\lambda_1 c_1 t_i^{(c_1-1)} e^{-\lambda_1 t_i^{c_1}} + (1 - \pi_1)\lambda_2 c_2 t_i^{(c_2-1)} e^{-\lambda_2 t_i^{c_2}}\right],$$

and the log likelihood functions for three segments Weibull model is

$$LL_3 = \sum_{i=1}^{n}\left[\pi_1\lambda_1 c_1 t_i^{(c_1-1)} e^{-\lambda_1 t_i^{c_1}} + \pi_2\lambda_2 c_2 t_i^{(c_2-1)} e^{-\lambda_2 t_i^{c_2}} + (1 - \pi_1 - \pi_2)\lambda_3 c_3 t_i^{(c_3-1)} e^{-\lambda_3 t_i^{c_3}}\right]$$

See Code Example 2 below for the maximum likelihood estimates of the three-latent class Weibull model.

Code Example 2.

```
proc nlp data=EEE.DATA_EEE_MOD maxfunc=50000 maxiter=50000;
      max loglik;
            parms   c1=2, c2=2, c3=2,
                lambda1 = 0.1, lambda2=0.1, lambda3=0.1,
                pi1=0.4, pi2=0.3;
            bounds  pi1 > 1e-12, pi1 <  1,  pi2 > 1e-12, pi2 <  1,
                lambda1 > 1e-12, lambda2 > 1e-12, lambda3 > 1e-12,
                        c1 > 0, c2 > 0, c3 > 0;
            lincon  pi1 + pi2 <  1;
      loglik = pi1*lambda1*c1*t**(c1-1)*exp(-lambda1*t**c1) +
pi2*lambda2*c2*t**(c2-1)*exp(-lambda2*t**c2) +
             (1 - pi1 - pi2)*lambda3*c3*t*(c3-1)*exp(-lambda3*t**c3) ;
run;
```

The parameter estimates were less stable than for Weibull-Gamma model in terms of dependence on initial values.

Expectations of Modeling Results vs. Reality

The Exponential-Gamma model exhibits negative duration dependence, or, in other words, it has a decreasing hazard function. In our case, this means that (given that a purchase did not occur before the time under consideration) conditional instantaneous rate of home purchasing goes down as time increases. Probably, this assumption is not realistic. Our intuition more or less gets along with opposite assumption: as the time of living in the US increases, the conditional instantaneous rate of purchasing a home is also increasing. Therefore, we can not expect that Exponential-Gamma model will be the best.

The Weibull-Gamma model is more flexible in terms of the hazard function behavior, and for the shape parameter $c > 1$, has upside-down “bathtub-shaped” hazard function. In other words, it exhibits positive and then negative duration dependence. In our case, this means that the conditional instantaneous rate of purchasing a home goes up during first years of the 20-year span under consideration, then it stabilizes, and after a certain point in time, it goes down. This assumption is in better agreement with our intuition.

In the analysis of duration data, the Weibull distribution is an appropriate distribution because its hazard function is flexible enough and is represented by a power function of time. So, depending on parameters, it could be monotonically decreasing or increasing. The latent class Weibull model, in general, could be a good candidate for modeling the data at hand, but since we know that heterogeneity is an essential factor for the data, the number of latent classes could be large enough. Here, we considered only two- and three-classes Weibull models due to the complexity of estimating the maximum likelihood. Another approach to parameter estimates (including the estimation of unknown number of classes) of general latent class Weibull model is described in (Marin et al., 2003). The flexibility of the hazard function for latent class Weibull model is incredible (Mosler and Sheicher, 2004), but the mixture of two or three distributions does not produce an upside-down “bathtub-shaped” hazard function. On the contrary, the hazard function has a regular bathtub-shaped hazard function. Therefore, we have expectations that a two or three class Weibull model would not be the best.

Results of modeling were more or less close to the expectations. The best was the class of Weibull-Gamma models with the following segmentation of the data:

Table 6.

Segment	Shape	Log-likelihood
Race = Asian	2.36	-781.96
Race = Other	2.25	-363.66
Race = White, Marital Status = Married	2.35	-792.34
Race = White, Marital Status = Non Married	2.22	-755.82

Conclusion

Instead of searching for the best segmentation, using some covariates, it makes sense to consider Weibull-Gamma model with covariates. In the absence of some essential covariates, taking into account the unobserved heterogeneity could significantly improve the quality of models.

References

Chapter 5. The NLP Procedure. SAS Institute Inc., *SAS/OR® User's Guide: Mathematical Program 8*, Cary, NC: SAS Institute, Inc., 1999, pp. 369-511.

Chapter 30. The Reliability Procedure. SAS Institute Inc., *SAS/QC® User's Guide, Volume 2, Version 8*, Cary, NC: SAS Institute, Inc., 1999, pp. 921-1028

Allison, Paul D. (1995), Survival Analysis Using the SAS System: A Practical Guide, Institute Inc., Cary, NC: SAS Institute, Inc., 1999.

Brusilovskiy, Eugene (2004), "The Piecewise Regression Model as a Response Modeling Tool", NESUG 2004 Conference Proceedings: http://www.nesug.org/html/Proceedings/nesug04/an/an09.pdf

Hardie, Bruce G.S. and Peter S. Fader, "Modeling Single Event Timing Data", Chapter 4 of unpublished working paper.

Hardie, Bruce G.S. and Peter S. Fader (2000), Applied Probability Models in Marketing Research http://brucehardie.com/talks/art00tut.pdf

Hardie, Bruce G.S. and Peter S. Fader (2001), Applied Probability Models in Marketing: Extension http://brucehardie.com/talks/art01tut_part2.pdf

Martin J.M., M.R. Rodrigez Bernal and M.P. Wiper (2003),Using Weibull Mixture Distributions to Model Heterogeneous Survival Data http://econpapers.hhs.se/paper/ctewsrepe/ws033208.htm

Mosler, Karl and Christoph Sheicher (2004), Homogeneity Testing in a Weibull Mixture Model http://www.uni-koeln.de/wiso-fak/wisostatsem/publications/HTiaWMM/WMIX.pdf

Author Contact Information:

Eugene Brusilovskiy
The Wharton School, University of Pennsylvania
3620 Locust Walk, Suite 1400
Philadelphia, PA 19104
Tel.: 215-898-2901
eugeneby@wharton.upenn.edu