Proportional Hazard Model: A Discussion with Serial Killer Data

Sayak Khatua

Abstract

In this paper I use the Weibull Proportional Hazard model to look at risks associated with the serial killer if they are not arrested, when their killing spree ends. I use a unique dataset on Serial Killers for the United States containing information on over 3000 reported serial killers in the history of the United States. I look at different groups of people based on age, sex and race and how the hazard and survival function change based on the specifications.

Contents

1	Introduction	2
2	Theoretical Foundation	3
2.1	Why not OLS ?	3
2.2	Proportional Hazard Models	4
2.3	Weibull Proportional Hazard Model	4
2.4	Log-logistic Proportional Hazard Model	5
2.5	Unobserved Heterogeneity (frailty)	6
2.6	Competing Risk Model	6
2.7	Estimation	7
3	Empirical Understanding	7
3.1	Data Description	7
3.2	Research Questions	9
3.3	Theoretical → Empirical	9
3.4	Results	9
3.4.1	Weibull Proportional Hazard Model	10
3.4.2	Weibull Proportional Hazard Model with unobserved heterogeneity . .	11
4	Conclusion	12
5	References	13

1 Introduction

Duration Analysis or Survival Analysis deals with the analysis of time duration until a specific event occurs. Duration analysis has been used in diverse fields such as engineering, economics, biostatistics, politics, etc. Duration data is generated from the need of analyzing likelihood of an event occurring when someone or something has been in a certain state. The idea is called 'failure time process'. A failure time process consists of units such as- governments, countries, individuals, etc. which are observed over a period of time, given they are all in some state, like individuals under trial, government without any corruption charges, countries at peace. At any given point there is a 'risk' of certain event occurring, where an event is the change from the current state. This point of change is called failure or censor. Few examples of duration analysis would be

1. Event: Return to Alcoholism

Time: Sobriety Period

Risk: Given that a person has stayed sober for 12 months, what is the likelihood of the person returning to drinking in the 13th month.

2. Event: Not Arrested

Time: Killing Spree

Risk: Given that a serial killer is on a killing spree 10 months, what is the likelihood that this person would return to killing in the 11th month if not arrested.¹

The estimation of duration models started receiving importance in the 1970s, when Cox(1972) proposed the use of Proportional Hazard Models in biostatistics and they were in no time adopted in economics. Duration analysis has been useful in the analysis of unemployment periods, recession periods, etc. Duration Model analysis basically has three main components, Survival, Failure and Risk. In this paper, I use duration data on serial killers, example (2) explains the idea of the analysis. I also discuss about different types of duration model and the best fit model for this analysis on serial killers. Woolridge(2010) talks about duration analysis, and explains the different models, and how to obtain the maximum likelihood

¹This description is based on notes from Brad Jones, August 2005, ICPSR MLE-2.

estimator. Hazard models are discussed Kiefer(1988), where he talks about the censoring of data and the different shapes of the hazard functions. The idea of duration analysis has been discussed in this paper based on popular literature and the data used to carry out the empirical analysis is on Serial Killers.

2 Theoretical Foundation

In this paper I am looking at only one kind of duration model, the Proportional Hazard Model. The proportional hazard model allows the inclusion of time invariant variables in the model whereas the baseline hazard model does not. There are a few reasons to why OLS is not used while analyzing the duration models and I discuss the reasons. Further, I discuss the different proportional hazard models, the weibull and the log-logistic models. Then the discussion goes into unobserved heterogeneity and finally into the competing risk model. The section ends with the discussion on estimation of hazard model with the maximum likelihood estimation.

2.1 Why not OLS ?

OLS can be run with duration data but there are certain aspects of duration, which cannot be captured by estimating with a simple OLS. For this reason there is a need to look into other models. Some of the problems are,

- **Normality Assumption**

In OLS regression analysis we assume that the duration data is normally distributed. But this assumption might not be true. In some cases if there are longer or shorter spells the data can right or left skewed. As a result of that the medians will describe the central tendency rather than the mean.

- **Negative Predicted Values**

Survival times are always positive and OLS can return negative predicted values which is theoretically not possible for durations.

- Censoring

OLS cannot distinguish between ‘censored’ and ‘uncensored’ observations. And dropping any of the two might create a selection bias and biased estimates.

- Time Varying Covariates ²

OLS cannot accomodate independent variables that change over time.

2.2 Proportional Hazard Models

A special class of models with time-invariant covariates, are called the proportional hazard function (Cox, 1972). The proportional hazard function includes covariates attached to baseline hazard function. The base line hazard is common to all, so the effects of the explanatory variables are always proportional. A generic formulation of the proportional hazard model is as follows:

$$\lambda(t; x) = k(x)\lambda_0(t) \quad (1)$$

where $k(.) > 0$ is a positive function of x and $\lambda_0(t)$ is the baseline hazard. There are two types of proportional hazard function, based on distribution of data, Weibull and Log-logistic proportional hazard function.

2.3 Weibull Proportional Hazard Model

The Weibull distribution is continuous probability distribution. It gives a distribution for which the failure rate is proportional to the power of time. It also allows for the hazard function to increase or decrease monotonically. The CDF and pdf of a generic Weibull distribution is as follows,

$$F(t) = 1 - e^{-\lambda t^\alpha} \quad \lambda > 0, \alpha > 0$$

²In this paper I only discuss models with time invariant covariates as the dataset used for empirical analysis contains only time invariant covariates

$$f(t) = \gamma \alpha t^{\alpha-1} e^{\gamma t^\alpha}$$

The hazard function is given by,

$$\lambda(t) = \gamma \alpha t^{\alpha-1} \quad (2)$$

The Weibull Proportional Hazard function is given by,

$$\lambda(t; x) = e^{x^\beta} \alpha t^{\alpha-1} \quad (3)$$

where $\gamma = e^{x^\beta}$, if $\alpha < 1$ failure rate decreases over time, $\alpha = 1$ failure rate is constant over time, $\alpha > 1$ failure rate increases over time.

2.4 Log-logistic Proportional Hazard Model

The Log-logistic distribution is a continuous probability distribution for a non-negative random variable. It also allows for the hazard function to increase or decrease monotonically. The CDF and pdf of a generic log-logistic distribution is given by;

$$F(x; \alpha, \beta) = \frac{x^\beta}{\alpha^\beta + x^\beta}$$

$$f(x; \alpha, \beta) = \frac{(\beta/\alpha)(x/\alpha)^{\beta-1}}{(1 + (x/\alpha)^\beta)^2}$$

The hazard function is given by,

$$\lambda(t) = \frac{\gamma \alpha t^{\alpha-1}}{1 + \gamma t^\alpha} \quad (4)$$

The log-logistic proportional hazard function is given by,

$$\lambda(t; x) = e^{x^\beta} \alpha t^{\alpha-1} [1 + e^{x^\beta} t^\alpha]^{-1} \quad (5)$$

where $\gamma = e^{x^\beta}$ if $\alpha < 1$ failure rate decreases over time, $\alpha = 1$ failure rate also decreases over time but starts from $t = 0$ with an y-intercept equal to γ , $\alpha > 1$ failure rate increases over time reaches maximum and then decreases, i.e. its bell shaped.

2.5 Unobserved Heterogeneity (frailty)

Generalization of duration models can be obtained by introducing unobserved heterogeneity in to the simple duration models. Lancaster(1979) argued that in specification of a hazard function there may be omitted explanatory variables, because some of the determinants of hazard function may be unobserved. In this paper, I introduce unobserved heterogeneity in the Weibull proportional Hazard model only for simplicity. The key assumptions used in this model are as follows,

1. The heterogeneity is independent of the observed covariates, as well as starting and censoring times.
2. The heterogeneity has a distribution known up to a finite number of parameters.
3. The heterogeneity enters the hazard function multiplicatively.

The Weibull proportional hazard function with unobserved heterogeneity is as follows,

$$\lambda(t, x, v) = e^{x\beta} \alpha t^{\alpha-1} v \quad (6)$$

where v is the heterogeneity with $E(v) = 1$. Lancaster(1979) assumed a gamma distribution for the v , and in this paper I assume the same just to be in line with some known result and for comparison with a known result. Gamma distribution is mainly used in conditional hazard which is shown in equation(6). But for unconditional hazard a Burr distribution is used, for the unobserved heterogeneity parameter.

2.6 Competing Risk Model

Generally, the duration models talk about either failing or not failing duration. But a competing risks model allows for more than two states, where the individual may exit into different alternatives. The hazard function becomes state specific i.e. for every exit alternative state there will be a hazard function. The different state models are estimated and their coefficients are compared. The competition risk model allows for comparison of specific failures rather than generalizing everything.

2.7 Estimation

The most common way of estimating a duration model is by using the maximum likelihood estimation. The models discussed above are all based on this estimation procedure. The estimation procedure discussed in this paper is a maximum likelihood estimation with censored data. Let, define a random variable, (a_i, t_i^*, x_i) where a_i is the time at which individual i enters the study, t_i^* denotes the length of time in initial state and x_i denotes the vector of observed covariates. As this is a censored data, it is assumed that, $t_i = \min(t_i^*, c_i)$, where c_i is the censoring time for individual i . Let d_i be an indicator whether the data is censored or not. Then the maximum likelihood estimators given by

$$\sum_{i=1}^N d_i \log f(t_i; \theta) + \sum_{i=1}^N (1 - d_i) \log[1 - F(t_i|x; \theta)] \quad (7)$$

The hazard function can be estimated once we have $\hat{\theta}$. This is the generalized maximum likelihood estimator, and it can be transformed to fit the different distributions. in this paper the Weibull is used for estimation, using the transformed MLE.

3 Empirical Understanding

In this paper I use the Weibull Proportional Hazard Model. The data used in this paper is on Serial Killers. I look at the duration serial killers are active over a period starting from 1960-2013. The data has been provided by Prof. Michael Aamodt, Radford University.

3.1 Data Description

The data contains information on 1698 serial killers over the period 1960-2013. The data on these serial killers obtained from Prof. Aamodt contains information on,

1. Sex
2. Race
3. Age at first kill

4. Victims
5. Country (Mostly, USA and Canada)
6. Date of first kill
7. Date of final kill
8. Date arrested
9. Period from first to last kill, i.e. the killing spree.

The dataset is modified to make it into a duration dataset. The killing spree data was firstly in years but was converted into monthly data. So, the dependent variable is the killing spree and it is measured in months. The data is right censored. The study ends when the serial killers get arrested and our data suggests there are serial killers who are at large still or there is no information about them. These observations have been censored. There are dummy variables generated for differentiating the white and non white population in this analysis. The event that I am looking at is that of the serial killer not being arrested. A dummy variable captures this event. The main idea of this study is to see if at all the serial killer is a risk after the end of his or her killing spree if they are not arrested.

The dataset is actually a flow dataset. A flow dataset in duration literature allows an individual to enter the state at any point in time during the interval of the study. The covariates are collected at the time the individuals enter the state. Flow dataset is generally right censored, i.e. after a certain amount of time information is not available for the individual for example after the serial killer gets arrested. In the serial killer dataset the spell of any serial killer can start at any time and end at time during the period of study. Figure: 1 of a representative sample of 5 individual serial killers to illustrate the flow nature of the data and the censoring done in the model. In Figure: 2 and Figure: 3 we see the mean, median and range of the killing spree between whites and non-whites and between certain age groups. Figure: 4 shows us that the data is skewed to the left. As a result an OLS estimation cannot be performed. Table 1 show the descriptive statistics of the dataset used in this paper.

3.2 Research Questions

1. Whether there is a risk of the serial killers to the population after they stop killing but are not arrested.
2. Do younger people have a longer span as serial killers?
3. Is there any difference in the killing spree based on race?
4. Do the number of victims increase as a serial killer clocks more kills?³

3.3 Theoretical → Empirical

The model I am using in this paper is the Weibull Proportional Hazard Model. The theoretical aspects of the model has been already explained in section 2.3. The empirical model has been shown here,

$$\lambda(t; x) = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5} \alpha t^{\alpha-1} \quad (8)$$

Figure 5, shows the baseline hazard model under a weibull distribution, and we can see from there that the α in the weibull proportional hazard model is going to be less than 1. From equation(8), x_i 's are victims, sex(dummy), race(white and nonwhite, dummy), killer not arrested(dummy), age at first kill. All the covariates, other than victims have been reported at the time the serial killer starts to kill or enters the period of study. The model is also estimated with the unobserved heterogeneity parameter, v . In this paper the unobserved heterogeneity parameter takes the form of a gamma distribution.

3.4 Results

The results show the weibull proportional hazard model in a general case and with unobserved heterogeneity. I report all of the results in the appendix section.

³Ideas developed mostly from Eric Hickey's Serial Murders and their Victims

3.4.1 Weibull Proportional Hazard Model

From Table 1, we find that the coefficient of the variable ‘Not Arrested’ is really high suggesting that there is good chance of the serial killer to return to killing. This estimate also tell us that there are other factors that determine the length of a killing spree. Police arresting or not arresting does not have much effect on the killing spree of the serial killer. This conclusion can be drawn because this dummy variable was created to suggest that arrests are the reason for serial killers to stop killing, but in reality its not. In my data all of the serial killers stop at a point irrespective of them being arrested or not.

The other estimated covariates suggest that higher victim count is associated with shorter spells. If the number of victims are high the most likely situation is that the serial killer has killed frequently in a short period of time. The coefficient for this variable is -0.139 . The sex dummy has a coefficient -1.039 which suggests that compared to a male serial killers female serial killers tend to have shorter killing spree. The data does not have a large number of female serial killers to male serial killers so the magnitude of the effect is difficult to judge. The negative sign with the coefficient suggests that the killing spree for women would be smaller. The last estimate is for the variable age at first kill, which is positive and suggests that younger serial killers have shorter killing spree than older serial killers. The coefficient for this variable is 0.013 which suggests that age does not have a huge significance on the killing spree of the serial killer.

The shape parameter, $\hat{\alpha}$ is less than 1 suggesting that the hazard function is decreasing at an increasing rate over time(Refer Figure 7). From the hazard ratios we can basically talk about the idea of duration dependence. By duration dependence I mean how the hazard ratio changes with time. All the variables except for ‘Not Arrested’ has a positive duration dependence meaning that the probability of a killing spree ending at time t (say) given the the killing spree has lasted to time t increases the duration. This suggests that the longer the killing spree, the probability of it being longer increases. The hazard ratio for ‘Not Arrested’ has not been shown in the table, but it is actually zero, showing that their is no duration dependence. This supports my earlier argument that the killing spree failure does not only depend on the person being arrested or not. There are other underlying unknown factors working here.

Figure 5 and Figure 6 show the baseline hazard and survival function without any covariates.

The Kaplan Meier survival estimates in Figure 8 suggest that at given point of time the probability of survival of killing spree for women is higher than that of men. Also the men's survival function has a higher slope than the women suggesting that men probability of survival decreases rapidly over time compared to women. The same exercise can be seen in Figure 9 where the survival estimate for any point in time suggest that the probability of white people to keep a killing spree going is higher than the nonwhite people. Although the slopes of the two functions are approximately same, suggesting that the decrease in probability of survival is approximately the same.

Figure 10 and Figure 11 show the smoothed hazard estimates. The first figure has separate estimates for men and women, and it suggests that the probability of failure increases over time and then reaches maximum and then starts decreases. But after the 300th month it becomes oscillatory only for the men though. The women's function is cut short because of lack of observations. Similarly, the next figure shows the smoothed hazard estimates for whites and non whites. For the whites the hazard function is smoother than the non-whites, which shows that over the span of the spells the hazard rate does not change much, but for the nonwhites the probability of failure increase as the span of spells increases.

3.4.2 Weibull Proportional Hazard Model with unobserved heterogeneity

The idea behind estimating a unobserved heterogeneity model was to account for the unobserved variables and also to check for spurious causation or endogeneity. After running the model with 100, 1000, 10000 iterations there is no convergence in this model. The frailty parameter θ is equal to zero and so it suggests that the model here does not actually have any sort of endogeneity at the individual observation level. The fact that no variance is reported for the frailty parameter, it is very difficult to interpret the results. In case the variance was zero the unobserved heterogeneity model would reduce to a normal weibull model.

4 Conclusion

The paper looks into the theoretical side of the proportional hazard models and uses Serial killer data to help explain the models further. I tried to account for unobserved heterogeneity in the models but the maximum likelihood estimator did not have a convergence. The model used here is the Weibull Proportional hazard model and the estimates show that serial killers after they stop killing i.e. at the end of their killing spree if they are not arrested they are not at all risky to the society. The hypothesis that arrests stop serial killers is not true and there are unobserved variables that affect the event of ending killing spree. It is seen that, compared to nonwhites, whites have a longer killing spree and their failure rate i.e. possibility of the killing spree ending stays stable over time. On the other hand women have a higher probability of the survival of the killing spree compared to men. Lastly, the paper shows that younger people have longer spells or killing spree and also whites have longer killing spree.

It would be interesting to see what happens to the estimates with new data on education, place of murders, weapon used, profession etc. This would be an interesting future avenue.

5 References

1. Cox, D. R. (1972): "Regression models and life tables (with discussion)", Journal of the Royal Statistical Society B, 34: 187-220.
2. Hickey, E. W. (2012). Serial murderers and their victims. Cengage Learning.
3. Kiefer, N. M. (1988). Economic duration data and hazard functions. Journal of economic literature, 646-679.
4. Lancaster, T. (1979): "Econometric Models for the Duration of Unemployment?", Econometrica, 47, 939-956.
5. Lancaster, T., The Econometric Analysis of Transition Data (Cambridge, 1990)
6. Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. MIT press.

Figure 1: Showing Evidence of Flow data

Figure 2: Box plots showing how Killing spree varies over age composition

Figure 3: Box Plots showing how Killing spree varies over race composition

Figure 4: Showing how the data is left skewed

Table 2: Weibull Proportional Hazard estimates

Variables	Parameter Estimates	Standard Error	Hazard Ratio
Victims	-0.139	0.035***	0.870
Sex(Female)	-1.039	0.376***	0.354
White	-0.325	0.108***	0.722
Not Arrested	17.862	0.055***	-
Age at first kill	0.013	0.007*	1.013
Constant	-21.087	0.261***	-
Shape Parameter	0.732	0.287	
Observations	1698		
Log-likelihood	-786.730		

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. The standard errors are robust.

Figure 5: Shows the Baseline hazard function

Figure 6: Shows the Survival Function under Weibull Proportional Hazard Model

Figure 7: Shows the Hazard Function under Weibull Proportional Hazard Model

Figure 8: Kaplan Meier Survival Estimates by Sex

Figure 9: Kaplan Meier Survival Estimates by Race-White and Non white

Figure 10: Smooth hazard estimates by Sex

Figure 11: Smooth hazard estimates by Race-White and Non white

Table 1: Shows descriptive statistics of Duration Data on Serial Killers

failure _d: **if_arrested**
analysis time _t: **durationinmonth**

Category	total	per subject			
		mean	min	median	max
no. of subjects	1452				
no. of records	1452	1	1	1	1
(first) entry time		0	0	0	0
(final) exit time		59.31198	1	19	557
subjects with gap	0				
time on gap if gap	0				
time at risk	86121	59.31198	1	19	557
failures	325	.2238292	0	0	1

failure _d: **if_arrested**
analysis time _t: **durationinmonth**

	time at risk	incidence rate	no. of subjects	Survival time		
				25%	50%	75%
total	86121	.0037738	1452	78	228	372