ISSN:1306-3111
e-Journal of New World Sciences Academy
2012, Volume: 7, Number: 1, Article Number: 3A0047

NWSA-PHYSICAL SCIENCES
Received: December 2011
Accepted: January 2012
Series : 3A
ISSN : 1308-7304

Ayhan Topçu
Fahrettin Arslan
Ankara University
ayhan_topcu@hotmail.com
Ankara-Turkey

COX ORANTILI HAZARD MODELİNİN PARAMETRİK MODELLERLE KARŞILAŞTIRMASI-SİMULASYON ÇALIŞMASI

ÖZET

Çalışma ile sağkalım analizinde yaygın olarak kullanılan ve yarı parametrik bir model olan Cox Regresyon modeli için veri üretilmesi incelenmiş ve Üstel/Weibull parametrik modelleri ile karşılaştırma amacıyla sansürlü ve sansürsüz durumlar için simulasyonlar yapılmıştır. Veri Üstel ya da Weibull dağılımından geldiğinde Cox modeli de uygulanabilir olmaktadır. Parametrik model tahminlerinin Cox modelinden elde edilen tahminlere göre daha doğru olması beklenmektedir. Burada bir karşılaştırma yapılmasının sebebi parametrik modeller yerine Cox modeli tercih edildiğinde ne kadarlık bir kayıp olduğunun görülebilmesidir.

Anahtar Kelimeler: Sağkalım Fonksiyonu, Hazard Fonksiyonu,
Cox Orantılı Hazard Regresyon Modeli,
Weibull Regresyon Modeli,
Üstel Regresyon Modeli

A COMPARATIVE SIMULATION STUDY BETWEEN COX PROPORTIONAL HAZARDS MODEL AND PARAMETRIC MODELS

ABSTRACT

In this study data generation for Cox Proportional Hazards model, a semi-parametric model commonly used in survival analysis, is investigated and a simulation study is performed in order to compare the parametric (Exponential - Weibull models) and semi-parametric model(Cox Model) for censored and uncensored data. Cox Model is applicable if data is coming from Weibull or Exponential distributions. The estimation results obtained from parametric models are expected to be more accurate than the results of the Cox models. In this paper, it is aimed to see how much accuracy is lost by choosing Cox proportional hazards model instead of parametric models.

Keywords: Survival Function, Hazard Function,
Cox Proportional Hazard Model,
Weibull Regression Model, Exponential Regression Model

1. GİRİŞ (INTRODUCTION)

Sağkalım analizinde başarısızlık olarak adlandırılan bir olay (hastalık-ölüm-bozulma-iflas) ortaya çıkana kadar geçen süre analiz edilir. Bu süre sağkalım (yaşam) süresi olarak adlandırılır. Sağkalım sürecini etkilediği düşünülen değişkenlere bağlı olarak sağkalım süresi modellenebilir. Bu modellemede eğer sağkalım sürelerinin dağılımı biliniyorsa parametrik modeller kullanılır. Dağılımın bilinmediği durumlarda 1972 yılında Cox tarafından öne sürülen ve yarı parametrik bir yöntem olan Cox regresyon modeli kullanılmaktadır. Cox Orantılı Hazard modelinde hazard fonksiyonlarının zamana orantılı olduğu varsayılır. Yani hazard oranı zamana karşı sabit, ya da bir bireyin hazard fonksiyonunun diğer bireyin hazard fonksiyonuna orantılı olması ve yaşam süresinden bağımsız olması anlamına gelmektedir. Cox regresyon modeli sağkalım sürelerinin dağılımına ilişkin varsayım gerektirmediğinden sağkalım analizinde oldukça yaygın olarak kullanılmaktadır.

T rasgele değişkeni sağkalım süresini gösteren ve $f(t)$ olasılık yoğunluk fonksiyonuna sahip bir rasgele değişken olmak üzere

$$S(t) = P(T > t) = \int_t^{\infty} f(x)dx \qquad 0 < t < \infty \tag{1.1}$$

olasılığına sağkalım fonksiyonu denir (Bireyin t zamanına kadar sağ olduğu bilindiğinde t den sonra sağkalma olasılığı).
Hazard fonksiyonu t zamanına kadar yaşadığı bilinen bir bireyin t zamanındaki ani ölüm riskini göstermektedir.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t} \tag{1.2}$$

şeklinde ifade edilir.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t, T \ge t)}{\Delta t} . \frac{1}{P(T > t)} = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t)}{\Delta t} . \frac{1}{P(T > t)}$$

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} \ln S(t) \tag{1.3}$$

olmaktadır.

Hazard fonksiyonu sağkalım analizinde dağılımı karakterize eden bir orandır. T rasgele değişkeninin sahip olduğu dağılıma göre hazard fonksiyonu farklı yapıdadır (Lawless, 1982).
t anı için hesaplanmış olan başarısızlık hızlarının birikimli fonksiyonu ise birikimli hazard foksiyonu olarak adlandırılır.

$$H(t) = \int_0^t h(u)du \tag{1.4}$$

$$S(t) = \exp\left[-H(t)\right] = \exp(-\int_0^t h(u)du) \tag{1.5}$$

Cox Orantılı hazard modelinde hazard fonksiyonları, açıklayıcı değişkenler $\underline{X}$ ve sağkalım süresi t nin bir fonksiyonu olarak her bir birey için aşağıdaki şekilde ifade edilir:

$$h(t,\underline{X}) = h_0(t)g(\underline{X}) \tag{1.6}$$

X'den bağımsız, zamandan bağımsız t'nin bir fonksiyonu

Burada $h_0(t)$ temel hazard fonksiyonu olarak adlandırılır. $\underline{X} = (X_1, X_2, ..., X_P)$ açıklayıcı değişkenler, $\underline{\beta} = (\beta_1, \beta_2, ..., \beta_P)$ açıklayıcı değişkenlerin katsayıları,

$g(\underline{X}) = \exp(\sum_{i=1}^{p} \beta_i X_i)$, açıklayıcı değişkenlerin logaritmik lineer formdaki fonksiyonudur.

Model lineer regresyon modeli olarak

$$\log h(t, \underline{X}) = \log(h_0(t) g(\underline{X})) = \log h_0(t) + \log g(\underline{X})$$

$$\log h(t, \underline{X}) = \log h_0(t) + \sum_{i=1}^{p} \beta_i X_i = \alpha(t) + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \quad (1.7)$$

şeklinde yazabilir.

Modelde $g(\underline{X})$ fonksiyonu üstel olduğundan hazard fonksiyonunun pozitifliği garanti edilmektedir. Ayrıca $X_1 = X_2 = ... = X_p = 0$ olması halinde hazard fonksiyonu;

$$h(t, \underline{X}) = h_0(t) \exp(0) = h_0(t) \quad (1.8)$$

olup temel hazard fonksiyonuna eşit olmaktadır. Bir başka açıdan bakıldığında, model açıklayıcı değişkenlerin temel hazard fonksiyonunu çarpan olarak ne ölçüde etkilediğini ortaya koymayı amaçlamaktadır.

$$S(t) = \exp(-\int_0^t h(u) du) \quad (1.9)$$

olduğundan orantılı hazard modeli için sağkalım fonksiyonu

$$S(t, \underline{X}) = \exp(-\int_0^t h(u, x) du) = \exp(-\int_0^t h_o(u) g(\underline{X}) du) \quad (1.10)$$

$$S(t, \underline{X}) = (\exp(-\int_0^t h_0(u) du))^{g(\underline{X})} = S_0(t)^{g(\underline{X})} \quad (1.11)$$

olmaktadır.

Modeldeki orantılı kavramı, bir bireye ait hazard'ın diğer bir bireyin hazardına oranının t'den bağımsız olmasıdır. $\underline{X} = (X_1, X_2, ..., X_p)$ ve $\underline{X}^* = (X_1^*, X_2^*, ..., X_p^*)$ iki bireye ait açıklayıcı değişkenler olmak üzere hazardların birbirine oranı;

$$\frac{h(t, \underline{X})}{h(t, \underline{X}^*)} = \frac{h_0(t) \exp(\sum_{i=1}^{p} \beta_i X_i)}{h_0(t) \exp(\sum_{i=1}^{p} \beta_i X_i^*)} = \exp \sum_{i=1}^{p} \beta_i (X_i - X_i^*) \quad (1.12)$$

olup, t'den bağımsız olmaktadır(Miller 1981:137).

Modelin bir başka özelliği $h_0(t)$ temel hazard fonksiyonuna ilişkin bir bilgi olmadan, $h_0(t)$' nin tahmini gerekmeden modelin üstel kısmından $\underline{\beta}$ parametresinin tahmin edilebilmesidir. Modelde dağılım bilinmemektedir, dolayısıyla $h_0(t)$ temel hazard fonksiyonu için belirli bir form yoktur. Bu özellik Cox modelini yarı - parametrik hale getirir (Kleinbaum 1996: 95).

2. ÇALIŞMANIN ÖNEMİ (RESEARCH SIGNIFICANCE)

Cox regresyon modeli için veri üretilmesi lineer regresyon modellerinden farklı olmaktadır. Lineer regresyon modellerinde bağımlı değişken belli bir dağılımdan gelen hata terimleri ve regresyon katsayılarının belirlenmesi ile üretilebilirken Cox Modelinde regresyon katsayılarının hazard fonksiyonu üzerindeki etkisi sağkalım sürelerine yansıtılacak şekilde sağkalım süreleri üretilmelidir. Bu çalışma popüler bir model olan Cox Orantılı Hazard Regresyon Modeli ve parametrik (weibull ve üstel) regresyon modelleri için sansürlü ve sansürsüz veri üretilmiş simulasyonlar yapılarak performansları karşılaştırılmıştır. Veri Üstel ya da Weibull dağılımından geldiğinde Cox modeli de uygulanabilir olmaktadır. Parametrik model tahminlerinin Cox modelinden elde edilen tahminlere göre daha doğru olması beklenmektedir. Çalışma ile oldukça yaygın olarak kullanılan Cox Modelinin Weibull ve Üstel parametrik modeller karşısındaki tahmin performansı incelenerek Cox modeli tercih edildiğinde ne kadarlık bir kayıp olduğunun görülmesi amaçlanmıştır.

3. METOT (METHOD)

Bu bölümde Cox modelinin parametrik regresyon modelleri ile karşılaştırılması amacıyla simulasyonlar yapılmıştır. Cox regresyon modeli için veri üretilmesi lineer regresyon modellerinden farklı olmaktadır. Lineer regresyon modellerinde bağımlı değişken açıklayıcı değişkenlerle ve hata terimleri ile doğrudan ilişkili olduğundan, bağımlı değişken belli bir dağılımdan gelen hata terimleri ve regresyon katsayılarının belirlenmesi ile üretilebilir. Ancak Cox Modeli hazard fonksiyonu ile belirlendiğinden, sağkalım süreleri üretilirken, regresyon katsayılarının hazard fonksiyonu üzerindeki etkisi sağkalım sürelerine yansıtılmalıdır (Bender et al. 2005:3).

Sağkalım analizinde fonksiyonlar arası ilişkilerden;

$$h(t)=\frac{f(t)}{S(t)}=-\frac{d}{dt}\log S(t) \qquad (3.1)$$

$$S(t)=\exp(-\int_0^t h(u)du) \qquad (3.2)$$

$$S(t)=\exp(-H(t)) \qquad (3.3)$$

elde edilir.

(2.1) ile verilen Cox Modeli için sağkalım fonksiyonu,

$$S(t,x)=\exp\left(-\int_0^t h_0(u)\exp(\beta x)du\right) \qquad (3.4)$$

$$S(t,x)=\exp\left(-H_0(t)\exp(\beta x)\right) \qquad (3.5)$$

olarak elde edilir.

Cox modeli için dağılım fonksiyonu;

$$F(t,x) = 1 - \exp\left(-H_0(t)\exp(\beta x)\right) \quad (3.6)$$

olmaktadır. Bir x rasgele değişkeninin dağılım fonksiyonu $F(x)$, (0,1) aralığında düzgün dağılıma sahip olduğundan,

$$U = 1 - \exp\left(-H_0(t)\exp(\beta x)\right) \sim U[0,1] \quad (3.7)$$

yazilabilmektedir.

Buradan Cox modeli için sağkalım süresi

$$T = H_0^{-1}\left[-\log(1-U)\exp(-\beta x)\right] \quad (3.8)$$

olarak elde edilir. (Bender et al. 2005:5).

Bu durumda Cox modeli için Üstel dağılımdan gelen sağkalım sürelerinin üretilmesi aşağıdaki şekilde olacaktır;

$$f(t) = \lambda\exp(-\lambda t),\quad \lambda > 0 \quad (3.9)$$

$$h_0(t) = \lambda \quad (3.10)$$

$$H_0(t) = \int_0^t \lambda dt = \lambda t \quad (3.11)$$

$$H_0^{-1}(t) = \frac{t}{\lambda} \quad (3.12)$$

$$T = \frac{\left[-\log(1-U)\exp(-\beta x)\right]}{\lambda} \quad (3.13)$$

λ, temel hazarda sahip Cox Modeli,

$$h(t,x) = \lambda\exp(\beta x) \quad (3.14)$$

şeklindedir.

4. BULGULAR (FINDINGS)

Veri Üstel ya da Weibull dağılımından geldiğinde Cox modeli de uygulanabilir olmaktadır. Parametrik model tahminlerinin Cox modelinden elde edilen tahminlere göre daha hassas olması beklenmektedir. Burada bir karşılaştırma yapılmasının sebebi parametrik modeller yerine Cox modeli tercih edildiğinde ne kadarlık bir kayıp olduğunun görülebilmesidir. Yapılan simulasyonlar için R programı kullanılmış, program çıktıları ekte verilmiştir.

Simulasyon hem sansürlü hemde sansürsüz durumlar için uygulanmıştır. n=20, 50, 100 olarak seçilmiş, döngü sayısı 100.000/n olarak alınmıştır. Sağkalım süreleri üstel dağılımdan üretilmiştir.

$\lambda(x) = \exp(2X)$, $X = 1/(1:n)$, T~Üstel(λ) olarak alınmış, $\beta = 2$ parametresi tahmin edilmiş, ortalama β tahminleri ve standart sapmalar her iki model için sağlanmıştır. Sonuçlar aşağıda sunulmaktadır.

Tablo 1. Simulasyon sonuçları-sansürsüz durum -weibull-cox regresyon modelleri karşılaştırması
(Table 1. Simulation result-uncensored data-weibull vs. cox regression models)

T~Üstel(λ) $\lambda(x)=\exp(2X)$ β=2 $X=1/(1:n)$	Weibull Regresyon Modeli		Cox Regresyon Modeli	
	Ortalama ($\hat{\beta}$)	Standart Sapma ($\hat{\beta}$)	Ortalama ($\hat{\beta}$)	Standart Sapma ($\hat{\beta}$)
n=20	2.3501	1.1773	2.4990	1.5460
n=50	2.2853	1.0034	2.3199	1.0914
n=100	2.2588	0.9466	2.2759	0.9859

Tablo 2. Simulasyon sonuçları-sansürlü durum -weibull-cox regresyon modelleri karşılaştırması
(Table 2. Simulation result-censored data-weibull vs. cox regression models)

Y~Üstel(λ) $\lambda(x)=\exp(2x)$ $\lambda=1$ β=2 $X=1/(1:n)$	Weibull Regresyon Modeli		Cox Regresyon Modeli	
	Ortalama ($\hat{\beta}$)	Standart Sapma ($\hat{\beta}$)	Ortalama ($\hat{\beta}$)	Standart Sapma ($\hat{\beta}$)
n=20	2.3630	1.1539	2.5337	1.5534
n=50	2.2910	1.0071	2.3396	1.1220
n=100	2.2907	0.9482	2.3253	0.9994

Tablo 3. Simulasyon sonuçları-sansürsüz durum -üstel-cox regresyon modelleri karşılaştırması
(Table 3. Simulation result-uncensored data-exponential vs. cox regression models)

Y~Üstel(λ) $\lambda(x)=\lambda\exp(2x)$ $\lambda=1$ β=2 $X=1/(1:n)$	Üstel Regresyon Modeli		Cox Regresyon Modeli	
	Ortalama ($\hat{\beta}$)	Standart Sapma ($\hat{\beta}$)	Ortalama ($\hat{\beta}$)	Standart Sapma ($\hat{\beta}$)
n=20	2.3302	1.1653	2.4985	1.5414
n=50	2.2442	1.0020	2.3089	1.1243
n=100	2.2955	0.9566	2.3096	1.0047

Tablo 4. Simulasyon sonuçları-sansürlü durum -üstel-cox regresyon modelleri karşılaştırması
(Table 4. Simulation result-censored data-exponential vs. cox regression models)

Y~Üstel(λ) $\lambda(x)=\exp(2x)$ $\lambda=1$ β=2 $X=1/(1:n)$	Üstel Regresyon Modeli		Cox Regresyon Modeli	
	Ortalam ($\hat{\beta}$)	Standart Sapma ($\hat{\beta}$)	Ortalama ($\hat{\beta}$)	Standart Sapma ($\hat{\beta}$)
n=20	2.3543	1.1799	2.5180	1.5263
n=50	2.3060	1.0562	2.3672	1.1766
n=100	2.2566	0.9603	2.8401	1.0091

5. SONUÇ (RESULT)

Sonuçlar incelendiğinde $\hat{\beta}$ ortalamalarının neredeyse aynı olduğu, Weibull ve Üstel modellerin standart sapmalarının Cox modeline göre biraz daha küçük olmakla beraber örneklem hacmi arttıkça birbirine çok yaklaştığı görülmektedir. Verinin sansürlü yada sansürsüz olması sonçlar üzerinde bir farklılık yaratmamıştır.
Sonuçlar her üç model uygulanabilir olduğunda Weibull ve Üstel parametrik modeller olmasına karşın yarı-parametrik bir model olan Cox modelinin de parametrik modellere oldukça yakın sonuçlar verdiğini göstermektedir. Bu durum parametrik modelleri uygulayama durumunda kullandığımız Cox Modelinin güvenilirliğini göstermektedir.

KAYNAKLAR (REFERENCES)

1. Bender, R., Augustin, T., and Blettner, M., (2005). Generating survival times to simulate Cox proportional hazards models, Statistics in Medicine 24, 1713-1723.
2. Cox, D.R., (1972). Regression models and life-tables, Imperial College, London, 187-220.
3. Kleinbaum, D.G., (1996). Survival Analysis a Self Learning Text. Springer, New York.
4. Lawless, J.F., (1982). Statistical models and methods for lifetime data, University of Waterloo, New Jersey.
5. Miller, R.G., (1981). Survival analysis, John Wiley& Sons.

EK (APPENDIX)

R Program Çıktıları

Weibull -Cox Karşılaştırması (sansürsüz durum)

```
> Regresyon<-function(n){
+ x<-1/(1:n)#aciklayici degisken
+ u<-runif(n,0,1)#normal dagilimdan uretilen veri
+ t<--(log(u))/exp(2*x) #ustel dagilimdan gelen sagkalim sureleri
+ p.model<-survreg(Surv(t,rep(1,length(x)))~x,dist="weibull")
+ cox<-coxph(Surv(t,rep(1,length(x)))~x)
+ return(c(abs(p.model$coef[2]),cox$coef))
+ }
> Simulasyon<-function(n){
+ sonuc<-matrix(NA,nrow=2,ncol=100000/n)
+ for(i in 1:(100000/n)) sonuc[,i]<-Regresyon(n)
+
return(c(mean(sonuc[1,]),mean(sonuc[2,]),sd(sonuc[1,]),sd(sonuc[2,])))
+ }
> Simulasyon(20)
[1] 2.350130 2.498974 1.177347 1.545999
> Simulasyon(50)
[1] 2.285274 2.319926 1.003437 1.091402
> Simulasyon(100)
[1] 2.2588211 2.2759405  0.9466347 0.9859422
```

Üstel -Cox Karşılaştırması (sansürsüz durum)

```
> Regresyon<-function(n){
+ x<-1/(1:n)#aciklayici degisken
+ u<-runif(n,0,1)#normal dagilimdan uretilen veri
+ t<--(log(u))/exp(2*x)#ustel dagilimdan gelen sagkalim sureleri
+ p.model<-survreg(Surv(t,rep(1,length(x)))~x,dist="exponential")
+ cox<-coxph(Surv(t,rep(1,length(x)))~x)
+ return(c(abs(p.model$coef[2]),cox$coef))
+ }
> Simulasyon<-function(n){
+ sonuc<-matrix(NA,nrow=2,ncol=100000/n)
+ for(i in 1:(100000/n)) sonuc[,i]<-Regresyon(n)
+
return(c(mean(sonuc[1,]),mean(sonuc[2,]),sd(sonuc[1,]),sd(sonuc[2,])))
+ }
> Simulasyon(20)
[1] 2.330155 2.498529 1.165334 1.541404
> Simulasyon(50)
[1] 2.2441804 2.308879 1.001971 1.124281
> Simulasyon(100)
[1] 2.2954798 2.3095979 0.9565959 1.0047299
```

Weibull -Cox Karşılaştırması (sansürlü durum)

```
> Regresyon<-function(n){
+ x<-1/(1:n)#aciklayici degisken
+ u<-runif(n,0,1)#normal dagilimdan uretilen veri
+ t<--(log(u))/exp(2*x)#ustel dagilimdan gelen sagkalim sureleri
+ u2<-runif(n,0,1) #normal dagilimdan uretilen veri
+ c<--(log(u))/exp(2*x) #ustel dagilimdan gelen sagkalim sureleri
+ obs<-pmin(t,c) #gozlemler
+ binary<-as.numeric(t<=c) #sansur verisi
+ p.model<-survreg(Surv(obs,binary)~x,dist="weibull")
+ cox<-coxph(Surv(obs,binary)~x)
+ return(c(abs(p.model$coef[2]),cox$coef))
+ }
> Simulasyon<-function(n){
+ sonuc<-matrix(NA,nrow=2,ncol=100000/n)
```

```
+ for(i in 1:(100000/n)) sonuc[,i]<-Regresyon(n)
+
return(c(mean(sonuc[1,]),mean(sonuc[2,]),sd(sonuc[1,]),sd(sonuc[2,])))
+ }
> Simulasyon(20)
[1] 2.363006 2.533693 1.153941 1.553409
> Simulasyon(50)
[1] 2.291024 2.339642 1.007054 1.122012
> Simulasyon(100)
[1] 2.290671 2.3252676 0.9481560 0.9994078
```

Üstel -Cox Karşılaştırması (sansürlü durum)

```
> Regresyon<-function(n){
+ x<-1/(1:n)#aciklayici degisken
+ u<-runif(n,0,1)#normal dagilimdan uretilen veri
+ t<--(log(u))/exp(2*x)#ustel dagilimdan gelen sagkalim sureleri
+ u2<-runif(n,0,1) #normal dagilimdan uretilen veri
+ c<--(log(u))/exp(2*x) #ustel dagilimdan gelen sagkalim sureleri
+ obs<-pmin(t,c) #gozlemler
+ binary<-as.numeric(t<=c) #sansur verisi
+ p.model<-survreg(Surv(obs,binary)~x,dist="exponential")
+ cox<-coxph(Surv(obs,binary)~x)
+ return(c(abs(p.model$coef[2]),cox$coef))
+ }
> Simulasyon<-function(n){
+ sonuc<-matrix(NA,nrow=2,ncol=100000/n)
+ for(i in 1:(100000/n)) sonuc[,i]<-Regresyon(n)
+
return(c(mean(sonuc[1,]),mean(sonuc[2,]),sd(sonuc[1,]),sd(sonuc[2,])))
+ }
> Simulasyon(20)
[1] 2.354260 2.518031 1.179938 1.526266
> Simulasyon(50)
[1] 2.305982 2.367190 1.056158 1.176629
> Simulasyon(100)
[1] 2.2566343  2.840052 0.9602973 1.0091227
```