www.istatistikciler.org
İstatistikçiler Dergisi 1 (2008) 16-22

İstatistikçiler Dergisi

COX REGRESYON MODELİ VE AKCİĞER KANSERİ VERİLERİ İLE BİR UYGULAMA

Durdu KARASOY

Hacettepe Üniversitesi Fen Fakültesi İstatistik Bölümü
06800-Beytepe, Ankara, Türkiye
durdu@hacettepe.edu.tr

ÖZET

Tıpta, salgın hastalıklara ve kronik hastalıklara ilişkin verilerin incelenmesi ve bu hastalıkları etkileyen faktörlerin saptanması için yaşam çözümlemesinde Cox regresyon modeli oldukça önemlidir. Bu çalışmada, Cox regresyon modeli tanıtılmış ve bu model kullanılarak akciğer kanserinde nüksü etkileyen risk faktörleri belirlenmiştir.

Anahtar Sözcükler: *Yaşam Çözümlemesi, Cox Regresyon Modeli, Hazard Fonksiyonu, Orantılı Hazard Modeli.*

ABSTRACT

**COX REGRESSION MODEL
AND AN APPLICATION TO LUNG CANCER DATA**

In medical science, in investigating the survival data of epidemic diseases and chronic diseases and determining the factors which affects these diseases, Cox regression model for survival analysis has gained widespread attention. In this paper, Cox regression model is presented and using this model risk factors affecting relapse of lung cancer is determined.

Key Words: *Survival Analysis, Cox Regression Model, Hazard Function, Proportional Hazard Model.*

1. GİRİŞ

20. yüzyılda başlayan yaşam çözümlemesi (survival analysis) çalışmaları, bu yüzyılın ikinci yarısı boyunca önemli ilerlemeler göstermiştir. 1972 yılında Cox tarafından geliştirilen regresyon modeli ile yaşam çözümlemesinde önemli adımlar atılmış, Cox'un yeni önerileri ve Kalbfleisch ve Prentice'nin katkıları ile bugünkü önemini kazanmıştır [5].

Bir birimin başarısızlığında zamanın etkisi olduğu kadar bazı özelliklerin de etkisi vardır. Araştırmacıların, bu özellikleri (değişkenleri) modele katma çalışmaları 1970'lere kadar pek yokken, Cox (1972)'un, Cox regresyon modeli ile ilgili makalesi çalışmalara yeni bir yön vermiştir.

2. YAŞAM ÇÖZÜMLEMESİ

Yaşayan bir organizmanın ya da cansız bir nesnenin belirli bir başlangıç zamanı ile ölümü (ya da belirlenen fonksiyonunu yerine getirememesi, başarısızlığı) arasında geçen zamana ‘‘yaşam

süresi'' ya da ''başarısızlık zamanı'' adı verilir. Herbir birey ya da birime ait yaşam süresi T, tanımı gereği sürekli ve pozitif bir değere sahiptir [4].

Birimin yaşam süresinin bir değerden büyük olma olasılığına yaşam fonksiyonu (survival function) adı verilir. Bu tanıma göre, f(t) yaşam süresinin olasılık yoğunluk fonksiyonu olmak üzere, yaşam fonksiyonu S(t),

$$S(t) = P(T \geq t) = \int_t^{\infty} f(u)du \ , \quad 0 < t < \infty$$

biçimindedir.

Bir birimin T = t zamanına kadar yaşaması koşulu altında, $\Delta t \rightarrow 0$ iken [t, t+Δt] aralığında yaşamının sona ermesi olasılığına hazard fonksiyonu denir. Hazard fonksiyonu h(t),

$$h(t) = f(t) / S(t) \ , \quad 0 < t < \infty$$

biçiminde elde edilir.

Hazard fonksiyonu, herbir zaman noktasındaki başarısızlık riskinin bir tanımıdır. Özellikle yaşam verilerinin modellenmesinde kullanılır. Hazard fonksiyonunu kullanmanın yararları aşağıdaki gibi sıralanabilir:
1. t yaşına kadar yaşamını sürdürdüğü bilinen bir bireyin var olan başarısızlık riskini dikkate alır.
2. Farklı grupların karşılaştırılması daha açık ve kesin olarak hazard fonksiyonu ile kolaylıkla yapılabilir.
3. Hazard fonksiyonunu temel alan modeller, durdurma ya da birden çok başarısızlık türü varken daha uygundur.
4. Başarısızlık zamanlarının üstel dağılıma sahip olduğu varsayılarak farklı grupların karşılaştırılması, özellikle hazard fonksiyonu kullanılarak yapıldığında daha kolaydır.
5. Hazard tek tür başarısızlık içeren sistemler için özel bir çözümleme biçimidir [3].

Yaşam verilerinin çözümlenmesi ile ilgili özel bir sorun, tüm birimlerin başarısızlık sürelerinin gözlenememesidir. Bazı birimler çalışmanın sonuna kadar yaşayabilir ya da birim herhangi bir sebeple çalışmadan çıkabilir. Bu durumda başarısızlık süresinin tam olmayan bir gözlemi, durdurulmuş (censored) olarak adlandırılır [1].

2.1. Yaşam Fonksiyonunun Tahmini

n birey üzerinden gözlenen ve $k \leq n$ olmak üzere başarısızlık süreleri $t_1 < ... < t_k$ biçiminde sıralandığında, t_j'deki başarısızlıkların sayısı d_j ile, yaşam süreleri gözlenmemiş bireyler için durdurma zamanları L_j ile gösterildiğinde S(t)'nin product - limit tahmini ya da Kaplan-Meier tahmini,

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) \qquad (1)$$

biçiminde tanımlanır. Burada n_j, t_j'de riskte olan bireylerin sayısıdır. Durdurma zamanı L_j gözlenen ölüm zamanına eşitse, durdurma zamanının ölüm zamanından ε kadar büyük olduğu

varsayılır. Bu varsayım, durdurulmuş bireye ait ölüm zamanının kendi durdurma zamanından daha büyük olacağı görüşünden kaynaklanmaktadır. Gruplar arasında yaşam olasılıkları açısından fark olup olmadığını incelemek için log-rank testi kullanılmaktadır.

Durdurma yoksa, $n_1 = n$ ve $n_j = n_{j-1} - d_{j-1}$ (j = 2, ... , k) olur ve (1) ifadesi, $\hat{S}(t) = \frac{\text{Gözlemlerin sayısı} \geq t}{n}$ ($t \geq 0$) biçiminde verilen deneysel yaşam fonksiyonuna indirgenir. Durdurma ve durdurmama durumlarının her ikisinde de $\hat{S}(t)$ bir adım fonksiyonudur ve t=0'da 1'e eşittir ve herbir yaşam süresi t_j'den hemen sonra $(n_j - d_j) / n_j$ kadar azalır [6, 7, 8].

3. COX REGRESYON MODELİ

Yaşam süresine ilişkin etkenlerin hazard fonksiyonu üzerindeki etkilerinin çarpımsal olduğu modeller, yaşam süresi verilerinin çözümlenmesinde önemli bir rol oynarlar. Bu modeller orantılı hazard modelleri olarak ifade edilir [7]. Orantılı hazard ailesi, farklı birimlerin hazard fonksiyonlarının birbirlerine orantılı olması özelliğine sahip modellerin bir sınıfıdır. Yani, x_1 ve x_2 regresyon vektörlerine sahip iki birimin hazard fonksiyonlarının oranı $h(t / x_1) / h(t / x_2)$ biçimindedir.

T bir birimin yaşam süresini temsil eden sürekli bir değişken ve x bu birimle ilgili bilinen açıklayıcı değişkenler vektörü olduğunda, T'nin hazard fonksiyonu,

$$h(t / x) = h_0(t)g(x)$$

biçimindedir. Buradaki g(x), birimlerin özelliklerini yansıtan x vektörünün hazard fonksiyonu üzerindeki çarpımsal etkisini belirleyen bir fonksiyondur. Çalışmalarda g(x) için değişik formlar görülmektedir. Örneğin, $g(x) = 1 + \beta^T x$ biçimindeki ifadeye doğrusal form, $g(x) = \log(1 + \exp(\beta^T x))$ biçimindeki ifadeye ise lojistik form adı verilir. Cox'un 1972'de incelediği model ise,

$$h(t / x) = h_0(t)\exp(x\beta)$$

biçimindedir [3]. Burada β regresyon katsayıları vektörü, $h_0(t)$ ise T'nin belirlenmemiş herhangi bir temel hazard fonksiyonudur. $h_0(t)$'ye "temel" denilmesinin nedeni, x=0 olması durumunda, $h(t/x) = h_0(t)$ olmasıdır. Bu model görünüm olarak regresyon modeline benzediğinden "Cox Regresyon Modeli" adı verilmektedir [3]. Bu modelde $g(x) = \exp(x\beta)$ biçiminde olmaktadır. Bu form g(x)'in en genel biçimidir ve uygulamalarda sıkça kullanılmaktadır. Çünkü, $\exp(x\beta)$ daima pozitiftir ve bu nedenle g(x)>0 koşulunun sağlanması için kısıt konulmasına gerek yoktur. $\exp(\beta)$ değerleri risk oranlarını göstermekte ve her bir düzeyin referans düzeye göre kaç kat daha fazla (β pozitifse) ya da kaç kat daha az (β negatifse) riskli olduğunu göstermektedir. Bu modelde $h_0(t)$ için özel bir biçim varsayılırsa "tam parametreli" bir Cox regresyon modeli elde edilir. Çok önemli bir model, temel hazard fonksiyonunun, $h_0(t) = h_p(ht)^{p-1}$ biçiminde olduğu, p=1 olduğunda üstel modelin elde edildiği Weibull modelidir [7, 9].

Cox regresyon modelinin önemli bir diğer avantajı ise, dağılımdan bağımsız olmasıdır. Diğer bir ifadeyle $h_0(t)$ için özel bir biçim varsayılmamıştır. Eğer veriler özel bir orantılı hazard modelinden geliyorsa, parametrik modele dayanan bir yaklaşım yerine dağılımdan bağımsız yaklaşımın kullanılması bir etkinlik yitimine neden olacaktır. Fakat bu etkinlik yitimi önemsenmeyecek kadar azdır [7].

Cox'un bu yaklaşımında, β için $h_0(t)$'ye bağlı olmayan bir olabilirlik fonksiyonu elde edilerek bu fonksiyon, β için bir kestirim elde edilebilecek biçimde maksimize edilmektedir [2].

4. UYGULAMA

Yaşam çözümlemesi uygulaması için akciğer kanserli 374 hasta verisi kullanılmıştır. Burada ilgilenilen olay, akciğer kanseri olan hastaların ameliyattan sonra hastalığının nüks etmesidir (201 hastada nüks olmuştur). Yaşam süresi, ameliyat tarihi ile hastalığın nüks etme tarihi arasında geçen zaman (ay) olarak alınmıştır. Hastalığı nüks etmeyenler için yaşam süresi durdurulmuş olarak kabul edilmiştir.

Önce Kaplan-Meier yöntemi ile yaşam olasılıkları tahmin edilmiş ve değişken düzeyleri arasında yaşam olasılıkları açısından fark olup olmadığını görmek için log-rank testi yapılmıştır. Daha sonra hastalığın nüks etmesini etkileyen faktörleri belirlemek için Cox regresyon çözümlemesi yapılmıştır.

Çözümleme için yaş, sigara tüketimi (paket-yıl), tümör boyutu, tümörün etrafa bulaşması (invazyon), patolojik evre değişkenleri kullanılmıştır.

Kullanılan değişkenler ve düzeyleri Çizelge 1'de verilmiştir.

Çizelge 1: Değişkenler ve düzeyleri

Değişken	**n**	**%**
Yaş		
≤39	20	5,3
40-49	65	17,4
50-59	119	31,8
60-69	130	34,8
≥70	40	10,7
Sigara tüketimi		
≤5	17	6,5
6-30	72	27,4
31-60	135	51,3
≥61	39	14,8
Tümör boyutu		
≤30	99	29,4
31-40	72	21,4
41-50	54	16,0
>50	112	33,2
İnvazyon		
0 (yok)	191	54,9
1 (var)	157	45,1
Patolojik evre		
Evre I	140	39,9
Evre II	94	26,8
Evre III	95	27,1
Evre IV	22	6,3

Kaplan-Meier yöntemi kullanılarak elde edilen yaşam olasılıkları Çizelge 2'de verilmiştir. Tüm veriler üzerinden genel üç yıllık yaşam olasılığı %93,7; beş yıllık yaşam olasılığı ise %88,2 olarak elde edilmiştir. Log-rank testi yapılarak tümör boyutu, invazyon ve patolojik evre

değişkenlerinin düzeyleri arasında yaşam olasılıkları açısından fark önemli bulunmuştur ($p<0,05$).

Dört düzeyli olan tümör boyutu değişkeninde, 1. ve 2. düzey için p=0,247, 1. ve 3. düzey için p=0,007, 1. ve 4. düzey için p=0,000, 2. ve 3. düzey için p=0,14, 2. ve 4. düzey için p=0,014, 3. ve 4. düzey için ise p=0,545 olarak elde edilmiştir. Bu değerlere bakıldığında 0,05'den küçük bulunanlar için o düzeyler arasındaki yaşam olasılığı farkının önemli olduğu yorumu yapılabilir.

Dört düzeyli patolojik evre değişkeninde de, 1. ve 2. düzey için p=0,001, 1. ve 3. düzey için p=0,000, 1. ve 4. düzey için p=0,000, 2. ve 3. düzey için p=0,607, 2. ve 4. düzey için p=0,003, 3. ve 4. düzey için ise p=0,003 olarak elde edilmiştir. Bu değerlere bakıldığında 0,05'den küçük bulunanlar için o düzeyler arasındaki yaşam olasılığı farkının önemli olduğu yorumu yapılabilir.

Çizelge 2: Kaplan-Meier sonuçları

Değişken	**Yaşam Olasılıkları (%)**	
	3 yıllık	**5 yıllık**
Genel	93,7	88,2
Yaş (0,651)		
≤39	0	0
40-49	60,4	34,8
50-59	73,9	59,5
60-69	73,7	61,4
≥70	38,3	0
Sigara tüketimi (0,298)		
≤5	52,0	52,0
6-30	70,4	46,1
31-60	75,5	65,7
≥61	37,9	37,9
Tümör boyutu (0,000)*		
≤30	72,8	61,2
31-40	67,0	40,9
41-50	50,1	0
>50	71,7	49,4
İnvazyon (0,002)*		
0 (yok)	84,6	71,6
1 (var)	81,4	64,4
Patolojik evre (0,000)*		
Evre I	79,8	73,7
Evre II	64,1	49,5
Evre III	61,6	48,6
Evre IV	0	0

Log-rank testi sonucunda hesaplanan p-değeri parantez içerisinde verilmiştir.
* p değeri 0.05'den küçük.

Bundan sonra Cox regresyon çözümlemesi öncelikle her bir değişken için ayrı ayrı yapılmış ve elde edilen sonuçlar Çizelge 3'de verilmiştir. Daha sonra ise beş bağımsız değişken birlikte ele alınarak geriye doğru seçim yöntemiyle nüks etmeyi etkileyen önemli faktörlerin yer aldığı bir model elde edilmiş ve bu model Çizelge 4'de verilmiştir.

Çizelge 3: Her bir değişken için Cox regresyon çözümlemesi

Değişken	β	Std. hata	p-değeri	Exp(β)	Güven aralığı (%95)
Yaş			0,661		
≤39				1	
40-49	-0,308	0,329	0,349	0,735	0,385 - 1,401
50-59	-0,114	0,303	0,708	0,893	0,493 - 1,617
60-69	-0,305	0,304	0,317	0,737	0,406 - 1,339
≥70	-0,085	0,348	0,806	0,918	0,465 - 1,815
Sigara Tüketimi			0,314		
≤5				1	
6-30	0,425	0,450	0,345	1,530	0,633 - 3,698
31-60	0,522	0,429	0,223	1,686	0,728 -3,905
≥61	0,804	0,467	0,085	2,234	0,895 – 5,579
Tümör boyutu			0,001*		
≤30				1	
31-40	0,242	0,228	0,288	1,274	0,815 – 1,992
41-50	0,631	0,242	0,009*	1,879	1,169 – 3,019
>50	0,746	0,195	0,000*	2,109	1,440 – 3,088
İnvazyon					
Yok				1	
Var	0,467	0,151	0,002*	1,595	1,187 - 2,142
Patalojik Evre			0,000*		
Evre I				1	
Evre II	0,644	0,194	0,001*	1,904	1,301 – 2,787
Evre III	0,778	0,187	0,000*	2,178	1,511 – 3,141
Evre IV	1,514	0,266	0,000*	4,546	2,701 – 7,653

* p değeri 0.05'den küçük.

Çizelge 3 incelendiğinde nüks etmeyi etkileyen faktörler, tümör boyutu, invazyon ve patolojik evre değişkenleri olarak bulunmuştur ($p<0,05$). Cox regresyon çözümlemesinde değişkenlerin ilk düzeyleri referans düzey olarak alınmıştır. Önemli bulunan değişkenlere ilişkin exp(β) değerleri yorumlandığında, tümör boyutu değişkeninin 41-50 düzeyinin referans olan ≤30 düzeyine göre 1,879 kat daha fazla riskli olduğunu (nüks olma açısından) söyleyebiliriz. Bu oran için güven aralığı ise 1,169 - 3,019 olarak bulunmuştur. Tümör boyutunun >50 düzeyi de yine önemli bulunan bir düzey olduğundan bu düzeyin de ≤30 düzeyine göre 2,109 kat daha fazla riskli olduğunu ve bu oran için de güven aralığının 1,44 – 3,088 olduğunu söyleyebiliriz. İnvazyon değişkeni için de invazyonu olanların, olmayanlara göre nüks olma riskinin 1,595 kat daha fazla olduğunu söyleyebiliriz. Bu risk için güven aralığı ise 1,187 – 2,142 olarak elde edilmiştir. Patalojik evre değişkeni için ise Evre II'nin Evre I'e göre 1,904 kat (1,301 – 2,787), Evre III'ün Evre I'e göre 2,178 kat (1,511 – 3,141), Evre IV'ün ise Evre I'e göre 4,546 kat (2,701 – 7,653) daha fazla riskli olduğunu söyleyebiliriz.

Çizelge 4 incelendiğinde beş değişken birlikte ele alındığında nüks etmeyi etkileyen faktörler olarak sadece invazyon ve patolojik evre değişkenleri bulunmuştur ($p<0,05$). İnvazyon değişkeni için invazyonu olanların, olmayanlara göre nüks olma riskinin 1,796 kat daha fazla olduğunu söyleyebiliriz. Bu risk için güven aralığı ise 1,142 – 2,823 olarak elde edilmiştir. Patalojik evre değişkeni için ise sadece Evre IV düzeyi önemli bulunmuş ve Evre IV'ün Evre I'e göre 4,981 kat (2,558 – 9,7) daha fazla riskli olduğunu söyleyebiliriz.

Çizelge 4: Geriye doğru seçim yöntemiyle Cox regresyon çözümlemesi

Değişken	β	Std. hata	p-değeri	Exp(β)	Güven aralığı (%95)
İnvazyon					
Yok				1	
Var	0,585	0,231	0,011*	1,796	1,142 - 2,823
Patalojik Evre					
Evre I			0,000*	1	
Evre II	0,194	0,288	0,501	1,214	0,691 - 2,133
Evre III	0,496	0,256	0,053	1,643	0,994 - 2,714
Evre IV	1,606	0,340	0,000*	4,981	2,558 - 9,700

* p değeri 0.05'den küçük.

5. SONUÇ

Bu çalışmada, yaşam çözümlemesinde önemli bir model olan Cox regresyon modeli tanıtılmış ve akciğer kanserinde nüksü etkileyen risk faktörleri belirlenmiştir.

Çalışma sonucunda, invazyon ve patalojik evre değişkenleri akciğer kanseri nüksünü etkileyen önemli değişkenler olarak bulunmuştur. İnvazyon değişkeni için invazyonu olanların, olmayanlara göre nüks görülme riskinin 1,796 kat daha fazla olduğu, patalojik evre değişkeni için ise Evre IV 'ün Evre I'e göre 4,981 kat daha fazla riskli olduğu sonucuna varılmıştır.

KAYNAKLAR

[1] Collate, D. (1994), *Modelling Survival Data in Medical Research*, Chapman and Hall, London.
[2] Cox, D. R. (1972), *Regression Models and Life Tables*, Journal of Royal Statistical Society, 34, 187-202.
[3] Cox, D. R. and Oakes, D. (1984), *Analysis of Survival Data*, Chapman and Hall, London.
[4] Johnson, R.E. and Johnson, N. (1980), *Survival models and data analysis*, John Wiley and Sons, New York.
[5] Kalbfleisch, J. D. and Prentice, R. L. (1973), *Marjinal Likelihoods Based on Cox's Regression and Life Model*, Biometrika, 60, 267-279.
[6] Kaplan, E. L. and Meier, P. (1958), *Nonparametric Estimation From Incomplete Observations*, Journal of the American Statistical Association, 53, 457-481.
[7] Lawless, J. F. (1982), *Statistical Models and Methods for Lifetime Data*, John Wiley and Sons, New York.
[8] Peterson, A.P. (1977), *Expressing the Kaplan-Meier estimator as a function of empirical survival functions*, J. Amer. Statist. Assoc., 72, 854-858.
[9] Prentice, R.L. (1973), *Exponential survival with censoring and explanatory variables*, Biometrika, 61, 539-544.