Anadolu Tarım Bilim. Derg., 2013,28(3):168-174
Anadolu J Agr Sci, 2013,28(3):168-174
doi: 10.7161/anajas.2013.28.3.168
URL: http://dx.doi.org/10.7161/anajas.2013.28.3.168

Derleme
Review

FARKLI VERİ YAPILARINDA KULLANILABİLECEK REGRESYON YÖNTEMLERİ

Arzu ARI Hasan ÖNDER*

Ondokuz Mayıs Üniversitesi, Ziraat Fakültesi, Zootekni Bölümü, Samsun
***honder@omu.edu.tr**

Geliş Tarihi : 23.01.2012 Kabul Tarihi : 23.11.2012

ÖZET: Regresyon, üzerinde durulan yanıt değişkeni ile ilişkili olabileceği düşünülen açıklayıcı değişkenlerin bir fonksiyonu olarak ifade edilmektedir. Değişkenler arasındaki ilişkinin fonksiyonel şekli regresyon modelleri ile incelenmektedir. Kullanılması gereken regresyon modeli verinin yapısına göre farklılık göstermekte ve yanlış model kullanılması hatalı sonuçların elde edilmesine neden olabilmektedir. Bu derlemede regresyon modellerinden; doğrusal regresyon, lojistik regresyon, negatif binom regresyon, poisson regresyon, temel bileşenler regresyonu, probit regresyon, ridge regresyon, Cox regresyon modellerinin hangi durumlarda kullanılabileceği incelenmiştir.
Anahtar kelimeler: Regresyon, Doğrusal regresyon, Doğrusal olmayan regresyonlar

REGRESSION MODELS USED FOR DIFFERENT DATA STRUCTURES

ABSTRACT: Regression can be expressed as a function between interested response variable and explanatory variables thought to be related on response. Functional form of the relationship between the explanatory variables and response variable described as regression model. The regression model must be chosen according to the data structure. If the chosen model is wrong, it leads to erroneous results. In this review, regression methods were examined to determine which regression models such as; linear regression, logistic regression, negative binomial regression, poisson regression, principal components regression, probit regression, ridge regression and Cox regression, is suitable for different data structure.
Key Words: Regression, Linear regression, Nonlinear regressions

1. GİRİŞ

Biyoloji, tıp, ekonomi, fizik, kimya ve sosyal bilimler gibi birçok alanda yaygın olarak kullanılmakta olan regresyon analizi, aralarında sebep - sonuç ilişkisi bulunan iki veya daha fazla değişken arasındaki ilişkiyi inceleyen ve bu ilişkiyi modellemek için kullanılan istatistiksel bir analiz yöntemdir (Vural, 2007). Uygun olmayan regresyon yöntemlerinin kullanılması hatalı ve yanıltıcı sonuçların elde edilmesine neden olabilmektedir. Regresyon analizinde incelenen değişkenler sürekli ya da kesikli yapıda olabilmektedir ve bu veri yapısına bağlı olarak farklı regresyon modelleri kullanılabilmektedir (Özarıcı, 1996).

Bu çalışma, regresyon yöntemlerinden; Doğrusal regresyon, Lojistik regresyon, Negatif binom regresyon, Poisson regresyon, Temel bileşenler regresyonu, Probit regresyon, Ridge regresyon, Cox regresyon yöntemlerinin hangi durumlarda kullanılabileceği konusunda araştırmacılara yol göstermek amacıyla yapılmıştır.

2. REGRESYON YÖNTEMLERİ

2.1. Doğrusal regresyon

Doğrusal regresyon analizi basit doğrusal regresyon ve çoklu doğrusal regresyon olarak iki başlık altında incelenmektedir.

Basit regresyon analizi, yanıt değişkeni ile tek bir açıklayıcı değişken arasındaki doğrusal ilişkiyi açıklar. Eğer tek bir yanıt değişkeni ve birden fazla açıklayıcı değişken arasındaki doğrusal veya eğrisel bir ilişki tanımlanmak istenirse, ilişki çoklu doğrusal regresyon analizi ile incelenir (Okur, 2009; Weisberg, 2005).

Basit doğrusal regresyonda, Y yanıt değişkeni, X_1 açıklayıcı değişkeni, β_0 ve β_1 bu değişkenin bilinmeyen parametrelerini ve ε_i şansa bağlı hata terimlerini ifade etmek üzere basit doğrusal regresyon modeli;

$$Y = \beta_0 + \beta_1 X_{1i} + \varepsilon_i \quad i = 1,2,...,n$$

olarak yazılabilmektedir (Kutner ve ark., 2005). Çoklu doğrusal regresyon modeli, p adet açıklayıcı değişken ve n adet gözlem için;

$$Y_i = \beta_{0+}\beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \varepsilon_i \quad i = 1, 2,..., n$$

şeklinde yazılabilmektedir (Kutner ve ark., 2005; Weisberg, 2005).

Gerek basit gerekse çoklu doğrusal regresyon analizi sonucunda elde edilecek olan regresyon modeline ait parametre kestirimlerinin güvenilir olabilmesi için modelle ilgili bazı varsayımların sağlanabilmesi gereklidir.

Basit doğrusal regresyon analizinde elde edilen regresyon denkleminin tahmin amaçlı kullanılabilmesi için; hata terimlerinin ($\epsilon_i = Y_i - \hat{Y}_i$) şansa bağlı normal dağılım göstermesi, hataların beklenen değerinin ortalamasının 0 ve varyansının homojen olup σ^2 'ye eşit olması, hataların bağımsız olması $[Cov(\epsilon_i , \epsilon_j)] = 0$, hata terimleri ile açıklayıcı değişken(ler) arasında korelasyon bulunamaması gibi bazı varsayımların sağlanması gerekmektedir (Alma ve Vupa, 2008).

Çoklu doğrusal regresyonda, basit doğrusal regresyondaki varsayımlara ilaveten açıklayıcı değişkenlerin birbirinden bağımsız olması varsayımının da sağlanması gerekmektedir (Vural, 2007). Açıklayıcı değişkenler arasındaki basit doğrusal korelasyon katsayılarının sıfır veya sıfıra çok yakın olması şartı seklinde de açıklanabilen bu varsayım, istatistikte "Çoklu doğrusal bağlantı" bulunmaması olarak ifade edilmektedir (Orhunbilge, 2002). Çoklu bağlantı durumda En Küçük Kareler (EKK) kestirim yöntemi işlevini yitirmektedir (Vural, 2007).

Bu nedenle açıklayıcı değişkenler seçilirken, bu değişkenlerin yanıt değişkeni ile basit doğrusal korelasyon katsayılarının yüksek (1'e yakın), birbirleri arasındaki basit doğrusal korelasyon katsayılarının düşük (0'a yakın) olmasına dikkat edilmesi önerilmektedir (Damodar, 2001). Bu varsayımların sağlanamadığı durumlarda parametre kestirim yöntemlerinin değiştirilmesi önerilmektedir.

2.2. Poisson regresyon

Araştırmadan elde edilen verilerin ölçeğinin sürekli yapıda olmadığı, diğer bir ifade ile kategorik veriye sahip olunduğunda doğrusal regresyon modelleri kullanılarak yapılacak analizler etkisiz, tutarsız ve güvenilmez sonuçlar verebilir. Özellikle, sayma ölçeğinde elde edilen veriler için kullanılabilecek en etkin modellerden biri Poisson regresyon modelidir (Deniz, 2005). Poisson regresyon modeli;

$$\log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots, \text{ olup,}$$
$$Y = (e^{\beta_0})(e^{\beta_1 X_1})(e^{\beta_2 X_2})\ldots,$$

olarak da tanımlanabilmektedir (Demaris, 2004). Modelden anlaşılacağı üzere Poisson regresyon modeli tahmin edicilerin doğrusal fonksiyonunun logaritmik dönüşümü olarak ifade edilebilmektedir (URL_1). İstatistik literatüründe bu model log-doğrusal model olarak bilinmektedir. $x\beta$ 'nın $\exp(x_i\beta)$ olarak alınması, beklenen sayma değerinin pozitif olmasını gerektirir. Bu durum Poisson dağılımı için zorunludur. Poisson dağılımında ortalama ve varyansın eşitliği söz konusu olup;

$$\mu_i = E(y_i|x_i) = V(y_i|x_i),$$

olarak gösterilebilir. Ortalama ve varyansın eşitliği "eşit yayılım" olarak ifade edilmektedir. Uygulamada sayılarak elde edilen değişkenler genellikle ortalamadan daha büyük varyansa sahiptirler. Bu durum aşırı yayılım (overdispersion) olarak adlandırılmaktadır. Aşırı yayılım durumunda Poisson Regresyon Modelinden elde edilen tahminler tutarlı fakat etkin değildir (Selim, 2003; Demaris, 2004).

2.3. Negatif Binom Regresyonu

Negatif binom regresyonunun uygulandığı veri kümesinde değerlerin çoğunun sıfır olmasından dolayı dağılış sağa çarpıktır. Bu durum, doğrusal regresyon kullanımında kuramsal olarak hatalı olan negatif parametre tahminlerini beraberinde getirmektedir (Frome ve ark., 1973; Cox, 1983; SAS, 2005). Negatif binom regresyon modeli için olasılık yoğunluk fonksiyonu;

$$\Pr(Y_i = y_i;\alpha,d) = \frac{(y_i + d - 1)!}{y_i!(d-1)!}\frac{\alpha^{y_i}}{(1-\alpha)^{y_i+d}}; \quad y_i = 0,1,2\ldots$$

olarak verilebilmektedir. Burada, α bir olgunun ortaya çıkma olasılığı $[P(Y=1)]$ ve d indeks veya yayılım parametresi olarak adlandırılmaktadır. Negatif binom regresyonu çözümlemesinde parametre kestirimleri Newton-Raphson algoritması yardımıyla En Çok Olabilirlik yöntemi kullanılarak elde edilir. Negatif Binom regresyonun model eşitliği;

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_q X_{iq} \text{ dir.}$$

Negatif Binom regresyonu, Poisson regresyonun özel bir durumudur. Bu iki model arasındaki seçim kararı, kestirimi elde edilen d katsayısının istatistiksel anlamlılığı yönünden belirlenir. Eğer d (yayılım parametresi), sıfırdan önemli derecede farklı değilse (istatistiksel olarak önemli değilse), Negatif Binom regresyonu Poisson regresyonuna dönüşür. Bununla beraber, d önemli derecede sıfırdan farklı ise, Negatif Binom regresyonu tercih edilmelidir (Hadayeghi, 2002; Aktaş ve Saraçbaşı, 2005).

2.4. Lojistik Regresyon

Lojistik regresyon, istatistikte kullanılan bir model oluşturma tekniği olup iki ya da daha fazla sınıfta ifade edilebilen kesikli verilerde yanıt değişkeni (Y) için bir model oluşturma tekniğidir. Yanıt değişkeninin kesikli olduğu durumlarda Lojistik ya da Probit regresyon yöntemleri kullanılmaktadır (Freese and Long, 2006). Modelin amacı, yanıt değişkeni iki değerli veya sınıflandırılmış olduğunda yanıt değişken ile açıklayıcı değişken veya değişkenler arasındaki ilişki açıklayan bir model oluşturmaktır (Önder ve Cebeci, 2002).

Lojistik regresyon ile doğrusal regresyon yöntemi arasında üç önemli fark vardır (Bircan, 2004), bu farklılıklar;

1. Doğrusal regresyon analizinde tahmin edilecek olan yanıt değişken sürekli iken, Lojistik Regresyon Analizinde yanıt değişken kesikli bir değer almaktadır.
2. Doğrusal regresyon analizinde yanıt değişkenin değeri, Lojistik Regresyon Analizinde ise yanıt değişkenin alabileceği değerlerden birinin gerçekleşme olasılığı tahmin edilir.
3. Doğrusal regresyon analizinde açıklayıcı değişkenin çoklu normal dağılım göstermesi şartı aranırken, Lojistik Regresyon Analizinde böyle bir şart yoktur.

Lojistik regresyon analizinde, yanıt değişkeni doğrudan modellenememektedir. Daha doğru bir yaklaşımla, lojistik regresyon analizi, Y yanıt değişkeninin değerinin birleştirilmiş olasılığı üzerine kurulmuştur. Uygulamada çok yaygın olarak yanıt değişkenin başarılı veya pozitif çıktı için 0 ve başarısız veya negatif çıktı için 1 değerini aldığı farz edilir. Yanıt değişken 1 olduğunda olasılık;

$$P(Y=1|X_1...X_p)=\frac{e^{\alpha+\sum_{j=1}^{p}\beta_j X_j}}{1+e^{\alpha+\sum_{j=1}^{p}\beta_j X_j}},$$

olarak gösterilebilir.

Lojistik regresyon modellerindeki çoklu iç ilişki, açıklayıcı değişkenler arasındaki güçlü korelasyondan kaynaklanır. Lojistik regresyondaki çoklu iç ilişki regresyon katsayılarının kestirimlerinin büyüklüğünün ve işaretinin yanlış olarak bulunmasını sağlayabilir ve neticede yanıt ve açıklayıcı değişkenler arasındaki ilişkiler hakkında yanlış sonuçlara ulaşılmasına yol açabilir (Ürük, 2007; Önder ve Cebeci, 2002; Kleinbaum ve ark., 1998).

2.5. Probit Regresyon

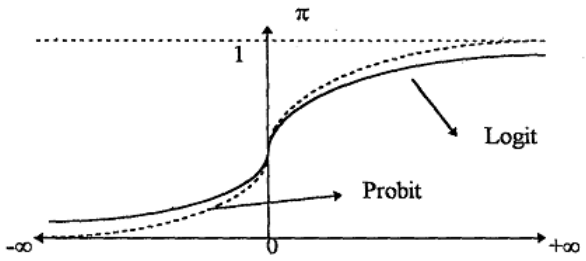
Probit analizi lojistik regresyona alternatif olarak bir veya daha fazla açıklayıcı değişkenin kategorik bir yanıt değişkeni (sağ, ölü; çalışıyor, çalışmıyor, ürün satıldı veya satılmadı vb) üzerindeki etkisini bulmak için kullanılan bir modeldir. Hem lojistik hem de Probit regresyon analizi birbirlerine oldukça benzer ve elde edilen olasılık tahminleri birbirlerine yakın değerdedir. Lojistik regresyon analizinde log-odds (olabilirlik oranları) kullanılırken, Probitte kümülatif normal dağılım kullanılmaktadır, Temel olarak Probit birikimli standart normal dağılışın tersidir (Topcu, 2008, Bek, 2009).

Probit modelin altında yatan varsayım, yanıt fonksiyonunun $Y_i = \alpha + \beta X_i + U_i$ formunda olmasıdır. Burada X_i gözlemlenebilen fakat Y_i gözlemlenemeyen değişkendir. $Y_i > 0$ ise $Y_i=1$, fakat $Y_i < 0$ ise $Y_i=0$ olur (Y_i değişkeninin sonucu atanırken, eşik değer olarak kullanılan c değeri genellikle 0 olarak alınmakta olup, sıfır yerine başka sayı değeri de kullanılabilmektedir (Demaris, 2004)). Eğer normal standart değişken Z için, $\Phi(z)$ kümülatif normal dağılım fonksiyonu $\Phi(z) = P(Z \leq z)$ olarak tanımlanırsa;

$$P(Y_i=1)=P(u_i>-\alpha-\beta X_i)=1-\Phi\left(\frac{-\alpha-\beta X_i}{\sigma}\right)$$

$$P(Y_i=0)=P(u_i\leq-\alpha-\beta X_i)=\Phi\left(\frac{-\alpha-\beta X_i}{\sigma}\right)$$

olarak ifade edilebilir. Probit modelinde birden fazla açıklayıcı değişken olduğu zaman, $\Pr(Y=1/X)=\Phi(X\beta)$ şeklinde tanımlanır. Burada Φ standart normal olasılık dağılımıdır. βX Probit skoru ya da indeksi olarak adlandırılır ve normal dağılıma sahiptir. Probit katsayısı β, tahmindeki bir birimlik artışın Probit skorunda yapacağı β standart sapmalık yükselmeyi ifade eder. Probit katsayısı bağımsız değişkenin bağımlı değişkene ait standart z-değerinde yapacağı etkiyi ölçer. Bu katsayıların sayısal büyüklüklerinin bir önemi ve özel bir yorumu yoktur, sadece ilişkinin yönü ve derecesini belirler (Topcu, 2008; Kulendran ve Wong, 2011; MacKinnon ve ark., 2007).



Şekil 1. Logit ve Probit birikimli dağılımlar

Probit ve Logit modeller genellikle yanıt değişkeninin iki düzeyli olduğu durumlarda uygulanmaktadır. Hesaplama zorluğu ve özel tablolara ihtiyaç duyulmaması bakımından Logit modelin Probit modeline göre kolaylığına rağmen, normal olasılık yoğunluk fonksiyonunun gerek teoride gerekse uygulamada daha çok kullanılan model olması biyolojik verilerin bazı özelliklerine ilişkin dağılımlarının normal olasılık yoğunluk fonksiyonuna uyum göstermesi, hesaplamada normal kümülatif dağılım fonksiyonunu kullanan Probit modelin uygulama çalışmasında kullanılmasını özendirmektedir (Özarıcı, 1996).

2.6. Temel bileşenler regresyonu

Her bir bağımsız değişkenin diğer bağımsız değişkenlerle arasındaki korelasyon katsayılarının karesi olan değer 1'e yakınsa, yüksek derecede çoklu bağlantı olduğu anlaşılabilir (Yıldırım, 2010). Açıklayıcı değişkenler arasında bir ya da daha fazla doğrusal bağıntı olduğunda çoklu bağlantı sorunu ortaya çıkmaktadır (Polat, 2009). Çoklu doğrusal bağlantı problemi, yapılan analizler sonucunda elde

edilen en küçük kareler kestiricilerinin varyans değerlerinin büyük olmasına ve tahminlerin gerçek değerlerinden uzaklaşmasına neden olmaktadır (Bulut ve Alın, 2009) ve çoklu bağlantı probleminin ortaya çıkması durumunda doğrusal regresyon analizi etkinliğini yitirmektedir. Bu durumda temel bileşenler regresyonu ya da ridge regreson yöntemi kullanılabilir.

Temel bileşenler regresyonu her doğrusal regresyon modelinin bir dik açıklayıcı değişkenler kümesine dayanarak yeniden açıklanması temeli üzerine yapılandırılmış olup, açıklayıcı değişkenler arasında çoklu bağlantı olduğu durumlarda uygulanmaktadır (Özkan, 2009).

Çoklu bağlantı durumunda, EKK yerine yanlı kestirim tekniklerinin kullanılmasının en uygun yaklaşım olduğu bilinmektedir (Albayrak, 2006).

Yanlı kestirimleri veren yöntemlerin başında, gerçek değişkenler yerine bunların dik dönüşümlerinin kullanıldığı Temel Bileşenler Regresyonu (PCR/Principal Component Regression) ve korelasyon matrisinin köşegen elemanlarına küçük bir pozitif sayı eklenerek kestirim varyanslarının küçültüldüğü Ridge Regresyon (RR) yöntemi gelmektedir (Polat, 2009).

Çoklu doğrusal regresyon modelinde açıklayıcı değişken katsayıları matris notasyonunda;

$$\hat{\beta} = (X'X)^{-1} X'y$$

olarak tahmin edilebilir ve burada, ***X*** açıklayıcı değişkenler matrisini ve ***Y*** ise yanıt değişkeni vektörünü göstermektedir., bu katsayılar temel bileşenler regresyonunda;

$$\hat{\beta}_{PC} = D_q \Lambda_q^{-1} D_q' X'y$$

olarak tahmin edilmektedir. Burada, D_q, $X'X$ ye ait ilk q adet öz vektör matrisi; Λ_q, $X'X$ ye ait ilk q adet öz vektör için köşegen matrisi göstermektedir (Al-Hassan ve Al-Kassab, 2009).

PCR sonucunda elde edilen tahminler yanlı olur. Ancak varyansın azaltılmasıyla, yanlılıktaki büyüklük dengelenir. RR tekniğinde k yanlılık sabitinin seçiminde yaşanan belirsizliğin aksine, PCR analizinde modelden çıkarılacak PC'lerin sayısı göreceli olarak daha tutarlıdır (Albayrak, 2011; Aswani ve ark., 2011; Al-Hassan ve Al-Kassab, 2009).

2.7. Rigde regresyon

Çoklu doğrusal bağlantı olduğunda yanlı tahmin yöntemlerinden Ridge Tahmin Yöntemi gerekli olan tüm değişkenlerin modele alınmasını sağlar. Bu yöntem çoklu doğrusal bağlantı olduğunda EKK kestirimlerinden daha küçük varyanslı parametre kesitirmlerinin elde edilmesini ve modelden gereksiz değişkenlerin çıkarılmasını amaçlamaktadır (Karadavut ve ark., 2005).

Ridge regresyon yöntemi genellikle modeldeki iki ya da daha fazla açıklayıcı değişken arasında yüksek dereceden ilişki olması durumunda kullanılır. Bu yöntemde uygulanırken ilk adım olarak açıklayıcı değişkenler standartlaştırılır (Karadavut ve ark., 2005). Standartlaştırılmamış orijinal değişkenlerin bulunduğu model;

$$y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i$$

i = 1, 2,…, n

şeklinde gösterilebilmektedir. Bu modeldeki açıklayıcı değişkenler standartlaştırılarak,

$$y_i = \mu + \gamma_1 Z_{i1} + \ldots + \gamma_p Z_{ip} + \varepsilon_i$$

modeli elde edilir (Karadavut ve ark., 2005). Ridge regresyonun yanlı regresyon yöntemi olmasına karşın EKK yöntemine göre iki önemli katkısı vardır. Bunlar; açıklayıcı değişkenlerde çoklu bağlantıyı elemine etmek ve regresyonda yanlılık karesiyle varyansı değiştirerek Hata Kareler Ortalamasını azaltmaktır (İpek, 2011).

Ridge regresyon modelinde açıklayıcı değişken katsayıları matris notasyonunda;

$$\hat{\beta}_{(k)} = (X'X + kI_p)^{-1} X'y,$$

olarak tahmin edilebilmektedir. Burada; k, yanlılık ya da Rigde parametresi olarak bilinmektedir ve veri kümesinden tahmin edilmelidir (Al-Hassan ve Al-Kassab, 2009).

Çoklu regresyon modelinde açıklayıcı değişkenler birbirleri ile bağlantılı olduklarında EKK β kestiricisinden daha küçük varyanslı β kestiricilerinden elde edilmesinde, güçlü çoklu bağlantı etkisi ile regresyon katsayılarında oluşan kararsızlıkların grafik üzerinde gösterilmesinde ve modeldeki gereksiz değişkenlerin çıkarılması amacıyla Ridge regresyon yöntemi kullanılabilmektedir (Yıldırım, 2010; Pamukçu ve ark., 2010).

2. 8. Cox regresyon

20. yüzyılda başlayan yaşam çözümlemesi çalışmaları, bu yüzyılın ikinci yarısı boyunca büyük ilerlemeler göstermiştir. Bu alandaki en etkili gelişmelerden biri; yaşam suresi üzerinde açıklayıcı değişkenlerin etkilerini ölçebilmek için kullanılan "Cox Regresyon Modeli" dir.

Parametrik modellerin gerektirdiği varsayımların (normallik, bağımsızlık vb.) sağlanmadığı durumlarda Cox regresyon analizi, parametrik analizlerden daha etkilidir. Cox regresyon modelinin temel varsayımları şu şekilde açıklanabilir: bağımsız değişkenlerin risk (hazard) fonksiyonu üzerindeki etkileri loglineerdir ve bağımsız değişkenlerin loglineer fonksiyonu ile risk fonksiyonu arasındaki ilişki çarpımsaldır (Özdamar,

2001; Laubender ve Bender, 2010; Chen ve ark., 2009; Liu ve ark., 2010).

Bu iki varsayıma ek olarak gözlemlerin birbirinden bağımsız olmaları ve risk oranının zamana göre değişmemesi, yani sabit olması gerekmektedir (Yay ve ark., 2007).

Cox regresyon modelinde yaşam zamanı T nin şartlı dağılımı, Z ve X hazard fonksiyonu $\lambda(.|Z,X)$ tarafından tanımlanmaktadır ve hazard;

$$\lambda(t|Z,X) = \lambda_0(t)\exp(\gamma Z + \beta' X) \ ,$$

olarak tanımlanabilmektedir. Burada; γ, Z ye ait regresyon katsayılarını, β ise X'e ait regresyon katsayıları vektörünü göstermektedir. $\lambda_0(.)$ ise Z=0 ve X=0 olan bir denek için temel alınan hazard fonksiyonunu göstermektedir (Laubender ve Bender, 2010).

Cox regresyon modelinin temel varsayımı olan orantılı hazard varsayımı, hazard oranının zamana karsı sabit olması ya da bir bireyin hazard fonksiyonun diğer bireyin hazard fonksiyonuna orantılı olması anlamına gelmektedir. Klinik denemelerde özellikle uzun süreli veriler söz konusu olduğun da orantısız hazardlar açığa çıkmaktadır. Hazardların orantılı olmaması durumunda ise Cox regresyon modeli yaşam verisi için uygun olmamaktadır (Ata ve ark., 2007; Chen ve ark., 2009).

3. MODEL SEÇİMİ

Yukarıda yapılan model tanımlarının ardından, model seçiminde uygulamaya yönelik karar mekanizması veriye ait histogram grafiğinin yapısına göre aşağıda örnek bir çalışma ile birlikte Çizelge 1'de verilmiştir.

4. SONUÇ

Bir araştırmada elde edilen verilere hangi istatistiklerin uygulanabilir olduğunu belirlemek için bazı ölçütler söz konusudur. Uygun istatistik yöntemlerle araştırmayı çözümlemek, araştırmanın güvenirliğini artırmakla birlikte sonuçların tutarlı bir şekilde yorumlanmasını da sağlamaktadır. Bu nedenle değişken yapıları, ölçme ölçekleri, varsayımların tutarlılığı istatistiksel çalışmalarda öncelikli olarak dikkate alınması gereken önemli bir durumdur. Çözümlemedeki adımsal yaklaşımı dikkate almadıklarından araştırmacılar yanıltıcı ve güvenilir olmayan raporlar yayınlayabilmektedir. Değişken yapıları seçilecek olan istatistiksel yöntemi belirlemektedir.

Elde edilen verinin yapısına uygun modelin seçilmesi araştırmadan elde edilen sonuçların güvenirliğini ve tahmin gücünü yükseltecektir.

5. KAYNAKLAR

Aktaş, A., Saraçbaşı,O., 2005. Negatif Binom Regresyon Modeli, VIII. Ulusal Biyoistatistik Kongresi, 20-22 Eylül 2005, Bursa, 124 – 129.

Albayrak, A. S. 2006. Uygulamalı Çok Değişkenli İstatistik Teknikleri, Asil Yayın Dağıtım Ltd. Şti.

Albayrak, A.S. 2011. Çoklu Doğrusal Bağlantı Halinde Enküçük Kareler Tekniğinin Alternatifi Yanlı Tahmin Teknikleri Ve Bir Uygulama. sbd.karaelmas.edu.tr/makaleler/1303-9245/200501001105126.pdf. Erişim Tarihi: 27.04.2011.

Al-Hassan, Y.M., Al-Kassab, M. M., 2009. A Monte Carlo Comparison between Ridge and Principal Components Regression Methods. Applied Mathematical Sciences, 3(42), 2085 – 2098.

Alma, G.Ö., Vupa, Ö., 2008. Regresyon Analizinde Kullanılan En Küçük Kareler Ve En Küçük Medyan Kareler Yöntemlerinin Karşılaştırılması. SDÜ Fen Edebiyat Fakültesi Fen Dergisi (E-Dergi). 2008, 3(2) 219-229.

Aswani, A., Bickel, P., Tomlin, C., 2011. Regression on Manifolds: Estimation of the Exterior Derivative. Ann. Statist. 39(1), 48 – 81.

Ata, N., Karasoy, D., Sözer, M.T. 2007. Orantısız Hazardlar İçin Tabakalandırılmış Cox Regresyon Modeli ve Meme Kanseri Hastaları Üzerine Bir Uygulama. Türkiye Klinikleri J Med Sci, s. 28.

Bek, Y., 2009. R Programında Doz-Yanıt Uygulamaları Çalıştayı Dinleyici Notları. 27 – 29 Mayıs 2009, Samsun.

Bircan, H. 2004. Lojistik Regresyon Analizi: Tıp Verileri Üzerine Bir Uygulama. Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 2004 / 2 : 185-208.

Bulut, E., Alın, A., 2009. Kısmi En Küçük Kareler Regresyon Yöntemi Algoritmalarından Nipals ve PLS - Kernel Algoritmalarının Karşılaştırılması ve Bir Uygulama. Dokuz Eylül Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 24(2), 127 – 138.

Chen, M., Ibrahim, J.G., Shao, Q., 2009. Maximum Likelihood İnference for the Cox Regression Model with Applications to Missing Covariates. Journal of Multivariate Analysis 100, 2018 – 2030.

Cox, R., 1983. Some Remarks on Overdispersion. Biometrika,70: 269-274.

Deniz, Ö. 2005. Poısson Regresyon Analizi. İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi Yıl:4 Sayı:7 Bahar 2005/1 S. 59-72.

Damodar, N. 2001. Temel Ekonometri, 2. Basım, İstanbul: Literatür Yayıncılık, s.192.

Demaris, A., 2004. Regression with Social Data : Modeling Continuous and Limited Response Variables. John Wiley & Sons, Inc. Hoboken, New Jersey.

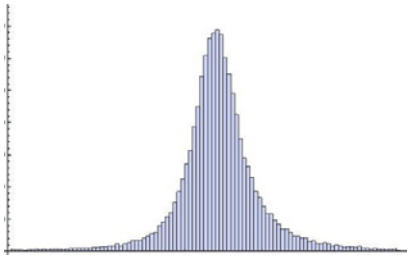
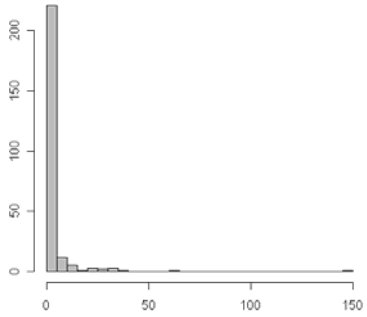
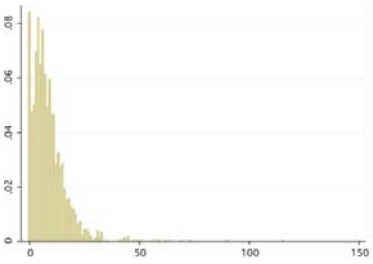
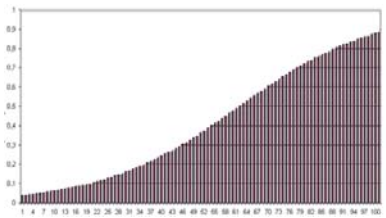
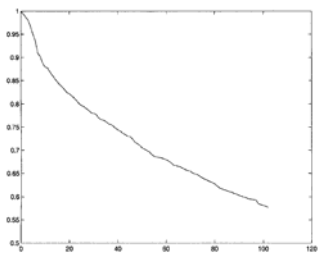
Freese, J. and Long, J.S., 2006. Regression Models for Categorical Dependent Variables Using. Stata. College Station: Stata Pres.

Frome, E.D., Kutner, M.H., Beauchamp, J.J., 1973. Regression Analysis of Poisson- Distributed Data. Journal of American Statistical Association, 68(344): 935-940.

Hadayeghi, A., 2002. Accident Prediction Models for Safety Evaluation of Urban Safety Transportation, Yüksek Lisans Tezi, Toronto Üniversitesi İnşaat Mühendisliği Bölümü, Kanada.

İpek, O. 2011. Rıdge Regresyon Üzerine Bir Çalışma. idari.cu.edu.tr/sempozyum/bil28.htm Erişim Tarihi: 02.05.2011.

Çizelge 1. Model seçimi için karar mekanizması

Yöntem	Histogram grafiği	
Doğrusal regresyon		Ölçme ölçeğinde ve EKK yönteminin varsayımları sağlanabiliyor ise kullanılmalıdır. Doğrusal regresyonun kullanılabileceği veri kümesine örnek olarak, yanıt değişkeni ve açıklayıcı değişkenlerin tamamının ölçülerek elde edildiği bir kasaplık canlı ağırlık (yanıt) üzerine doğum ağırlığı ve sütten kesim ağırlığının (açıklayıcı) etkisi verilebilir (URL_2)
Poisson		Sayma ölçeğinde ve *d* parametresi sıfırdan istatistiksel olarak farklı değil ise kullanılmalıdır. Poisson regresyonun kullanılabileceği veri kümesine örnek olarak, klinik mastitis düzeyinin (skor) yanıt değişkeni olduğu ve açıklayıcı değişkenin somatik hücre sayısı olduğu bir deneme gösterilebilir (URL_3)
Negatif binom		Sayma ölçeğinde ve *d* parametresi sıfırdan istatistiksel olarak farklı ise kullanılmalıdır. Negatif Binom regresyonun kullanılabileceği veri kümesine örnek olarak, böcek sayısı (yanıt) üzerine bitki çeşidi (indeks) ve bitkide yaprak sayısının (açıklayıcı) etkisinin incelendiği bir araştırma verilebilir (URL_4)
Lojistik ve Probit		İki veya daha fazla sınıfta gruplanan verilerde, birikimli artış ifadelerinde Probit kullanılmalı, olabilirlik tahmini için ise Lojistik kullanılmalıdır. Lojistik regresyonun kullanılabileceği veri kümesine örnek olarak, hayvan sağlığı (sağlıklı, hasta) yanıt değişkeni üzerine yaş (kesikli), süt verimi (sürekli), parazit (var, yok) ve barınak durumunun (skor) açıklayıcı değişkenlerinin etkisinin incelendiği bir çalışma gösterilebilir. Probit regresyonun kullanılabileceği veri kümesine örnek olarak, orman yangını (var, yok) yanıt değişkeni üzerine nem (sürekli), organik madde tabakasının kalınlığı (sürekli), organik madde tabakasının çeşidi (indeks), sıcaklık (sürekli) açıklayıcı değişkenlerinin etkisinin incelendiği bir çalışma verilebilir (URL_5).
Temel Bileşenler ve Ridge	Doğrusal regresyonda çoklu bağlantı olması durumunda, ilgisiz değişkenlerin modelden çıkarılması için Ridge kullanılmalı, tüm değişkenler modelde tutulmak isteniyor ise Temel Bileşenler kullanılmalıdır. Temel bileşenler ve Ridge regresyonun kullanılabileceği veri kümesine örnek olarak, çoklu bağlantı olması durumunda, canlı ağırlık (sürekli) yanıt değişkeni üzerine vücut ölçülerinin (sürekli) etkisinin incelendiği bir çalışma gösterilebilir.	
Cox		Uzun süreli yaşam verilerinin modellenmesinde kullanılmalıdır. Cox regresyonun kullanılabileceği veri kümesine örnek olarak, sığırlarda subklinik endometritis riski üzerine ilk tohumlamada gebelik oranı (yüzde), gebelik başına servis sayısı (oran), parity (doğum sırası; kesikli), vücut kondisyon skoru (skor) açıklayıcı değişkenlerinin etkisinin incelendiği bir çalışma gösterilebilir (Demaris, 2004)

Karadavut, U., Genç, A., Tozluca, A., Kınacı, İ., Aksoyak, Ş., Palta, Ç., Pekgör, A. 2005. Nohut (*Cicer Arietinum* L.) Bitkisinde Verime Etki Eden Bazı Karakterlerin Alternatif Regresyon Yöntemleriyle Karşılaştırılması. Tarım Bilimleri Dergisi 2005, 11 (3) 328-333.

Kleinbaum, D. G., Kupper, L. L., Muller, K. E., Nizam, A., 1998, Applied Regression Analysis and Other Multivariable Methods, 798, Duxbury Press, 511 Forest Lodge Road Pacific Grove, CA 93950 USA.

Kulendran, N., Wong, K.K.F., 2011. Determinants versus Composite Leading Indicators in Predicting Turning Points in Growth Cycle. Journal of Travel Research, 50(4), 417 – 430.

Kutner, M. H., Nachtsheim, C. J., Neter, J. ve Li, W., 2005. Applied Linear Statistical Models. McGraw-Hill Irwin Companies inc. New York.

Laubender, R.P., Bender, R., 2010. Estimating Adjusted Risk Difference (RD) and Number Needed to Treat (NNT) Measures in the Cox Regression Model. Statist. Med. 29, 851 – 859.

Liu, M., Lu, W., Shore, R.E., Zeleniuch-Jacquotte, A., 2010. Cox Regression Model with Time-Varying Coefficients in Nested Case–Control Studies. Biostatistics, 11 (4), 693 – 706.

MacKinnon, D.P., Lockwood, C.M., Brown, C.H., Wang, W., Hoffman, J.M., 2007. The Intermediate Endpoint Effect in Logistic and Probit Regression. Clinical Trials, 4, 499 – 513.

Okur, S. 2009. Parametrik Ve Parametrik Olmayan Doğrusal Regresyon Analiz Yöntemlerinin Karşılaştırmalı Olarak İncelenmesi. Çukurova Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tez, Adana.

Orhunbilge, N. 2002. Uygulamalı Regresyon Ve Korelasyon Analizi, İ.Ü. İşletme Fakültesi Yayınları, İstanbul.

Önder, H., Cebeci, Z., 2002. Lojistik Regresyonlarda Değişken Seçimi. Çukurova Üniv. Ziraat Fakültesi Dergisi, 17(2),105–114.

Özarıcı, Ö. 1996. Farklı Not Sistemlerinde Öğrencinin Başarılı Olma Olasılığının Probit Regresyon Analiziyle Değerlendirilmesi, Osmangazi Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi.

Özdamar, K. 2001.SPSS ile Biyoistatistik, Kaan Kitabevi, Eskişehir.

Özkan, K. 2009. Toprağın Tarla Kapasitesi Değişiminin Toprak Türüne Göre Temel Bileşenler Regresyon Analizi İle Modellenmesi Süleyman Demirel Üniversitesi Orman Fakültesi Dergisi Seri: A, Sayı: 2, Issn: 1302-7085, Sayfa: 1-9.

Pamukçu, E., Çolak, C., Çalık, S., Kuzu, Z., 2010. Sistolik Kan Basıncının Tahmininde Yanlı Regresyon Yöntemlerinin Kullanılması. Journal of Inonu University Medical Faculty, 17(4), 347 – 353.

Polat, E. 2009. Kısmi En Küçük Kareler Regresyonu. Hacettepe Üniversitesi, Fen bilimleri Enstitüsü, Yüksek Lisans Tezi.

SAS, 2005. SAS/STAT Software:Hangen and Enhanced. SAS, Inst. Inc., USA.

Selim, S. 2003. Sayma Veri Modelleri İle Çocuk Sayısı Belirleyicileri: Türkiye'deki Seçilmiş İller İçin Sosyoekonomik Analizler. D.E.Ü.İ.İ.B.F.Dergisi Cilt:18 Sayı:2, Ss:13-31.

Topcu, Y. 2008. Çiftçilerin Tarımsal Destekleme Politikalarından Faydalanma İstekliliğinde Etkili Faktörlerin Analizi: Erzurum İli Örneği. Akdeniz Üniversitesi Ziraat Fakültesi Dergisi, 21(2), 205–212.

URL_1: http://www.oxfordjournals.org/our_journals/tropej/online/ma_chap13.pdf. Poisson Regression Analysis (Erişim Tarihi: 01.05.2011).

URL_2: http://theclimatescepticsparty.blogspot.com/2011/08/nzclimate-truth-newsletter-no-273.html (Erişim Tarihi: 10.01.2012).

URL_3: http://www.ats.ucla.edu/stat/r/dae/zipoisson.htm (Erişim Tarihi: 10.01.2012).

URL_4: http://www.philender.com/courses/categorical/notes1/trunc0.html (Erişim Tarihi: 10.01.2012).

URL_5: http://archimede.bibl.ulaval.ca/archimede/fichiers/23662/ch07.html (Erişim Tarihi: 10.01.2012).

Ürük, E. , 2007. İstatistiksel Uygulamalarda Lojistik Regresyon Analizi. Marmara Vural, A. 2007. Aykırı Değerlerin Regresyon Modellerine Etkileri ve Sağlam Kestiriciler. Marmara Üniversitesi Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi, İstanbul.

Vural, A. 2007. Aykırı Değerlerin Regresyon Modellerine Etkileri ve Sağlam Kestiriciler. Marmara Üniversitesi Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi, İstanbul.

Weisberg, S., 2005. Applied Linear Regression. John Wiley & Sons, Inc. Hoboken, New Jersey.

Yay, M., Çöker, E., Uysal, Ö. 2007. Yaflam Analizinde Cox Regresyon Modeli ve Artıkların İncelenmesi. Cerrahpaşa Tıp Dergisi, 38:139-145, ISSN: 1300-5227.

Yıldırım, N., 2010. En Küçük Kareler, Ridge Regresyon Ve Robust Regresyon Yöntemlerinde Analiz Sonuçlarına Aykırı Değerlerin Etkilerinin Belirlenmesi. Çukurova Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tez, Adana.