

Logistic Regression Analysis Based on Jackknife Method: An Application on the Estimation of Coronary Artery Disease

Hayriye Esra Akyuz
Faculty of Science and Arts, Bitlis Eren University, Bitlis, 13000 Turkey
E-mail: heakyuz@beu.edu.tr

Abstract

Coronary artery disease is a multifactorial disease characterized by various factors and their interactions, and is the most common cardiac disease. In this study, it is aimed to obtain the logistic regression model for the estimate of coronary artery disease and to compare the parameter estimates for the factors affecting the estimation of parameters to the parameter estimates for the logistic regression model based on the jackknife method. The parameter estimates are made by backward elimination method and a suitable regression model is determined. In the data analysis and parameter estimates, the R 3.3.3. and SPSS 23.0 package software programs are used. It is obtained that the classification percentages of the model are over 80%. When the results of the original logistic regression analysis and the logistic regression analysis based on the jackknife method are examined, it is seen that there is little difference between the coefficient estimates. It has been determined that the standard errors of the parameter estimates in the Jackknife method are lower than the standard errors of the parameter estimates obtained from the classical sample. As a result; it was found that the parameter estimates based on the Jackknife method are very effective and the logistic regression model based on this method was found to be quite successful for the estimation of coronary artery disease.

Keywords: Backward elimination, Jackknife, Coronary artery disease.

Jackknife Yöntemine Dayalı Lojistik Regresyon Analizi: Koroner Arter Hastalığının Tahmini Üzerine Bir Uygulama

Özet

Koroner arter hastalığı, çeşitli faktörler ve bunların etkileşimi tarafından belirlenen bir multifaktöriyel hastalıktır ve en sık görülen kalp hastalığıdır. Bu çalışmada koroner arter hastalığının tahmini için lojistik regresyon modelinin elde edilmesi ve koroner arter hastalığını etkileyen faktörlere ait parametre tahminlerinin, jackknife yöntemine dayalı lojistik regresyon modelinin parametre tahminleri ile karşılaştırılması amaçlanmıştır. Parametre tahminleri geriye doğru eleme yöntemi ile yapılmış ve uygun regresyon modeli belirlenmiştir. Veri analizlerinde ve parametre tahminlerinde R 3.3.3. ve SPSS 23.0 paket programları kullanılmıştır. Modelin sınıflama yüzdelerinin %80'nin üzerinde olduğu elde edilmiştir. Orjinal lojistik regresyon analizi sonuçları ile jackknife yöntemine dayalı lojistik regresyon analizi sonuçları incelendiğinde ise katsayı tahminleri arasında çok az bir farklılık olduğu görülmüştür. Jackknife yönteminde parametre tahminlerinin standart hatalarının, klasik ornekten elde edilen parametre tahminlerinin standart hatalarından, daha düşük olduğu belirlenmiştir. Sonuç olarak; jackknife yöntemine dayalı parametre tahminlerinin oldukça etkin olduğu elde edilmiş ve bu yöntemine dayalı lojistik regresyon modelinin, koroner arter hastalığını tahmin etmede oldukça başarılı olduğu görülmüştür.

Anahtar Kelimeler: Geriye doğru eliminasyon, Jackknife, Koroner arter hastalığı.

1. Giriş

Koroner arter hastalığı (KOA), çeşitli genler, çevresel faktörler ve bunların etkileşimi tarafından belirlenen kompleks bir multifaktöriyel hastalıktır ve en sık görülen kalp hastalığıdır (Gündoğdu vd. 2007, Yin vd. 2012, Elsaid vd. 2014). Bu sebeple Koroner Arter Hastalığını etkileyen faktörlerin istatistiksel olarak belirlenmesi amaçlanmıştır. Bu amaç doğrultusunda Lojistik regresyon analizinden faydalanılmıştır.

Regresyon yöntemleri bir bağımlı değişkenle bir ya da daha fazla bağımsız değişken arasındaki ilişkiyi incelemek için kullanılır. Bu yöntemler arasında en bilinenleri basit ve çoklu doğrusal regresyon yöntemleridir. Bu yöntemlerde bağımlı değişken sayısal veri tipindedir. Buna karşılık çalışmalarda bağımlı değişkenin kategorik/nitel veri tipinde olması durumu ile de sıklıkla karşılaşılır. Bu durumda parametre tahminlerini yapmak için En Küçük Kareler yönteminin kullanılması uygun olmamaktadır. Böyle durumlarda lojistik regresyon analizi kullanılabilir. Nitel bağımlı değişkenin kategori sayısına kategorilerin nominal ya da ordinal olmasına göre farklı lojistik regresyon modelleri mevcuttur (Alpar. 2011). Lojistik regresyon, bağımsız değişken veya değişkenlerin bağımlı değişken üzerinde yapmış olduğu etkileri, bağımlı değişkenin iki kategorisinden birinin gerçekleşme olasılığının diğer kategorinin gerçekleşme olasılığıyla karşılaştırılmasından faydalanarak analiz eder (Kleinbaum ve Klein. 2002).

Lojistik modelin biyolojik deneylerin analizi için kullanımı ilk olarak Berkson (1944) tarafından önerilmiş. Coşkun vd. (2004) Lojistik Regresyon analizinin dış hekimliğinde bir uygulamasını gösterirken, Vupa ve Çelikoğlu (2006) akciğer kanseri hastalar için lojistik regresyon modeli önermişlerdir. Koroner arter hastalığının tahmin edilmesinde lojistik regresyon analizi çeşitli çalışmalarda kullanılmıştır (Hirashiki et al., 2003; Horibe et. al., 2004; Çolak et. al., 2007).

Jackknife örnekleme yöntemi güven aralıkları ve dağılımların elde edilmesinde bir istatistiğin standart hatasını tahmin etmeye olanak sağlar. Bazı durumlarda regresyon analizinin varsayımları sağlanmayabilir. Böyle durumlarda jackknife yöntemi kullanılabilir. Bu çalışmada koroner arter hastalığının tahmini için lojistik regresyon analizine dayalı uygun modelin elde edilmesi ve bu model ile elde edilen parametre tahminlerinin jackknife yöntemine dayalı olarak elde edilen parametre tahminleri ile karşılaştırılması amaçlanmıştır.

İzleyen bölümde lojistik regresyon analizi ve jackknife yöntemi üzerinde durulmuştur.

2. Lojistik Regresyon Analizi

Lojistik regresyonda bağımlı/yanıt değişkenin değerinin kestirilmesi ile ilgilenilmez. Bunun yerine bağımlı değişkenin 1 değerini alması olasılığı (riskli durum 1 olarak belirlendiğinde) tahmin edilmeye çalışılır. Elde edilen sonuç 0 ile 1 arasında değerler alabilir. Modelde iki ya da daha fazla bağımsız değişken olması durumunda çoklu lojistik regresyon yönteminden yararlanır. İlgilenilen değişkenlerin sayısı p olmak üzere bağımsız değişkenler vektörü $X = (X_1, X_2, \dots, X_p)$ ile gösterilsin. Buradan çoklu lojistik regresyon modeli;

$$\pi(x) = P(Y = 1 / x) = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}} \quad (1)$$

biçiminde verilir (Alpar, 2011). Bu model bağımlı değişkenin odds'u türünden aşağıdaki gibi belirtilebilir.

$$\frac{\pi(x)}{1 - \pi(x)} = e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)} \quad (2)$$

Odds'un doğal logaritması alınarak lojit dönüşüm ise aşağıdaki gibi yapılmış olur.

$$\text{lojit} \pi(x) = \ln \left(\frac{\pi(x)}{1 - \pi(x)} \right) \quad (3)$$

Odds'un doğal logaritması alındığında model doğrusal modele dönüşür.

$$g(x) = \ln \left(\frac{\pi(x)}{1 - \pi(x)} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \quad (4)$$

Çoklu lojistik regresyon analizinde en az değişken ile en iyi uyuma sahip model elde edilmek istenir. Bu amaçla veriye en iyi uyan model belirlenmek için ileriye yönelik ekleme (forward) ve geriye doğru eleme (backward) gibi adımsal yöntemlerden yararlanır. İleriye yönelik seçim yönteminde sadece sabitin olduğu bağımsız değişkenlerin olmadığı model ile işleme başlanır. Modele eklendiğinde log olabilirlik üzerinde en fazla değişime neden olan bağımsız değişken belirlenir ve bu değişken modele alınır. Eklenen değişkenin modele katkısı önemsiz bulunana kadar modele bağımsız değişken eklenmeye devam edilir. Geriye doğru eleme yönteminde ise ilk aşamada tüm değişkenler modele alınır. Her adımdan modelden çıkartıldığında sapmada en küçük artışa sebep olan değişken modelden çıkarılır.

2.1. Jackknife Yöntemi

Jackknife yeniden örnekleme yöntemi bir istatistiğin yanlılık miktarını ve standart hatasını tahmin etme olanağı sağlar. İlk olarak Quenouille (1956) tarafından bir istatistiğin yanlılığını ortadan kaldırmak için kullanılmıştır. Daha sonra Tukey (1958) tarafından hipotez testi ve güven aralıkları için kullanılmıştır. Örnekleme, yığından örneklerin seçilmesi işlemidir. Bir örneklemin en önemli özelliği; yığın hakkında geçerli ve tutarlı tahminlere ulaşabilmek için en küçük örnekleme hatasını elde edebilmesidir.

Bununla birlikte, parametrik tahmin yöntemlerinin, örnek sayısı küçük olduğunda ve aynı zamanda parametrik test varsayımları sağlanmadığında güvenilir sonuçlar vermedikleri bilinir. Böyle durumlarda yeniden örnekleme yöntemleri kullanılabilir. Bunlarda biri Jackknife Yöntemidir. Jackknife yöntemi, yığın parametrelerinin tahmininde güven aralıkları elde ederek örnekleme hatasını en aza indirecek şekilde geliştirilmiştir.

X_1, X_2, \dots, X_n rasgele örneği için $\hat{\theta} = S(x)$ tahmin edicisi için ilgilendiğimizi düşünelim. Jackknife yöntemine göre i . gözlem çıkarıldığında yeni örnek aşağıdaki gibi elde edilir.

$$x_{(i)} = (x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n); i = 1, 2, \dots, n$$

Yeni örneğe dayalı tahmin edici ise

$$\hat{\theta}_{(i)} = S(x_{(i)})$$

biçiminde elde edilir. Bu durumda yanlılık miktarı aşağıdaki gibi olur:

$$\text{yanlılık} = (\hat{\theta}_{(i)} - \hat{\theta})$$

Burada $\hat{\theta}_{(i)}$, θ parametresinin tahmini olup aşağıdaki gibi elde edilir.

$$\hat{\theta}_{(i)} = \frac{\sum_{i=1}^n \hat{\theta}_{(i)}}{n}$$

Standart hatanın Jackknife tahmini;

$$sh_{jack} = \frac{\left\{ \frac{\sum_{i=1}^n (\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2}{n-1} \right\}^{1/2}}{\sqrt{n}}$$

olarak verilir (Fenwick, 1979; Abdi ve Williams, 2010).

2.2. Jackknife Delete-d Algoritma

Lojistik regresyon analizinde jackknife yönteminin uygulanması aşamasında, mevcut bağımlı ve bağımsız değişkenlerden her defasında bir gözlem hariç tutularak, en çok olabilirlik yöntemi uygulanır. Bu süreç örnek hacmi kez tekrar edilir. Daha sonra, model parametrelerinin tahminine ait pseudo değerler, model standart hatası, modelin belirleme katsayısı ve korelasyonu hesaplanır. Jackknife yöntemine dayalı regresyon modeli, hesaplanan değerlerin ortalaması alınarak tahmin edilir. Bu algoritmanın adımları aşağıdaki gibidir (Efron, 1982; Şahinler and Topuz, 2007):

Adım 1: Yığından n hacimli bir X_1, X_2, \dots, X_n rasgele örnek seçilir ve örnek d büyüklüğünde S bağımsız gruba bölünür.

Adım 2: Tüm örnekten ilk d gözlem seti çıkarılır ve (n-d) boyutunda kalan gözlem kümesinden Maksimum Olabilirlik yöntemi ile logistic regresyon katsayıları $(\hat{\beta}^{j2})$ tahmin edilir.

Adım 3: İkinci d gözlem çıkarılır ve (n-d) boyutunda kalan gözlem kümesinden en çok olabilirlik yöntemi ile logistic regresyon katsayıları $(\hat{\beta}^{j2})$ tahmin edilir.

Adım 4: n hacimli örnekten her bir d gözlem çıkarılır ve katsayılar $\hat{\beta}^{jk}$ biçiminde gösterilerek tahmin edilir. Burada $\hat{\beta}^{jk}$ jackknife regresyon katsayısı vektörüdür.

Adım 5: $\hat{\beta}^{j1}, \hat{\beta}^{j2}, \dots, \hat{\beta}^{js}$ Jackknife tahminlerinin olasılık dağılımı $F(\hat{\beta}^j)$ elde edilir.

Adım 6: Jackknife regresyon katsayılarının tahmini aşağıdaki gibi elde edilir.

$$\hat{\beta}^j = \frac{\sum_{k=1}^n \hat{\beta}^{(jk)}}{s} = \bar{\beta}^{(jk)}.$$

3. Materyal ve Metot

Bu çalışmada parametre tahminleri geriye doğru eleme yöntemi ile yapılmıştır ve uygun regresyon modeli belirlenmiştir. Parametre tahminlerinde En Küçük Kareler metodu kullanılmıştır. Hata terimlerinin deneysel dağılımı $\hat{\beta}$ tahminlerinden elde edilmiştir. Veriler Bitlis Kamu Hastaneleri Genel Birliğinden gerekli izinler alınarak, 2017 yılında Bitlis Devlet Hastanesi Kardiyoloji Bölümüne başvuran 170 Koroner Arter hastası ve 170 sağlam olmak üzere 340 kişiden elde edilmiştir. Veriler üzerine lojistik regresyon analizi uygulanarak, hastaların koroner arter hastalığı riski taşıyıp taşımadıklarına ilişkin açıklayıcı bir model oluşturulmaya çalışılmıştır. Veri analizlerinde ve parametre tahminlerinde R 3.3.3. ve SPSS 23.0 istatistik paket programları kullanılmıştır. Koroner Arter hastalığının tahmininde kullanılan bağımsız değişkenler yaş, hemoglobin, ürik asit, yüksek dansiteli lipoprotein (HDL), düşük dansiteli lipoprotein (LDL), toplam bilirubin, direkt bilirubin olarak seçilmiştir ve en uygun değişkenler geriye doğru eleme yöntemi ile belirlenmiştir. Bağımlı değişken olarak koroner arter hastalığı alınmıştır. Daha sonra aynı veri seti üzerinde jackknife yöntemi uygulanarak parametre tahminleri yapılmış ve parametre tahminleri elde edilmiştir. En son adım olarak

da lojistik regresyon sonucu elde edilen parametrelerin tahmin deęerleri ile jackknife yöntemi sonucu elde edilen parametrelerin tahmin deęerleri karşılaştırılmıştır.

Üzerinde çalışılan yığınım tüm gözlemlerini parametre tahmini için kullanmak hem zaman kaybına yol açacak hem de maliyeti arttıracaktır. Yeniden örnekleme yöntemlerinden biri olan jackknife yönteminin bilgisayar destekli olarak kullanımı bu sorunları ortadan kaldırıp büyük avantajlar sağlamaktadır. Hastaların koroner arter hastalığı riski taşıyıp taşımama riski üzerine yapılan bu çalışmada, jackknife yönteminden elde edilen sonuçların lojistik regresyon analizi sonucunda elde edilen parametre tahminlerinin standart hataları ile karşılaştırılarak daha güvenilir parametre tahminlere ulaşmak için etkili olup olmadığı incelenmiştir. Çalışmada $Y_i=1$ olması hastada koroner arter varlığını, $Y_i=0$ ise yokluğunu ifade etmektedir.

4. Bulgular

Tablo 1. Koroner Arter hastalığının tahmininde kullanılan bağımsız deęişkenlere ait bazı tanımlayıcı deęişkenler

Deęişkenler	Minimum	Maksimum	Ortalama	Standart sapma
Yaş	38.00	78.00	55.7471	8.3260
Hemoglobin	9.00	25.50	13.7312	2.0239
Ürik asit	2.60	9.00	5.0182	1.3239
HDL	19.00	58.00	37.2353	8.0780
LDL	61.00	202.00	127.4765	25.0321
Toplam bilirubin	0.10	1.30	0.7619	0.2700
Direkt bilirubin	0.01	0.50	0.1785	0.0927

Tablo 1’de Koroner Arter hastalığının tahmininde kullanılan bağımsız deęişkenlere ait deęerlere yer verilmiştir. Bu sonuçlara göre çalışmada kullanılan 340 koroner hastasının yaş aralığının 38 ile 78 arasında deęiştığı ve yaş ortalamasının 55 olduğu görülmüştür. Hastaların ortalama hemoglobin deęerinin 13.73 ± 2.02 , ortalama ürik asit deęerinin 5.01 ± 1.32 , ortalama HDL deęerinin 37.23 ± 8.08 , ortalama LDL deęerinin 127.47 ± 25.03 , ortalama toplam bilirubin deęerinin 0.76 ± 0.27 ve ortalama direk bilirubin deęerinin ise 0.17 ± 0.09 olduğu belirlenmiştir.

Tablo 2. Lojistik regresyon modeline dahil edilecek değişkenlerin belirlenmesi ve analiz sonuçları

Adım	Değişken	Katsayı	Standart hata	Wald istatistiği	p-değeri	Odds oranı	% 95 güven aralığı	
							Alt sınır	Üst sınır
1	Yaş	0.150	0.023	44.013	<0.001	1.162	1.112	1.215
	Hemoglobin	-0.070	0.087	0.639	0.424	0.933	0.786	1.106
	Ürik asit	0.393	0.137	8.236	0.004	1.481	1.133	1.937
	HDL	-0.006	0.021	0.084	0.771	0.994	0.954	1.036
	LDL	0.063	0.009	50.410	<0.001	1.065	1.047	1.084
	Tbilirubin	2.661	0.721	13.630	<0.001	14.313	3.485	58.787
	Dbilirubin	-0.670	1.779	0.142	0.707	0.512	0.016	16.712
	Sabit	-19.177	2.431	62.243	<0.001	0.000		
2	Yaş	0.150	0.023	43.817	<0.001	1.162	1.111	1.215
	Hemoglobin	-0.063	0.085	0.561	0.454	0.939	0.795	1.108
	Ürik asit	0.382	0.131	8.481	0.004	1.465	1.133	1.894
	LDL	0.063	0.009	50.581	<0.001	1.065	1.047	1.084
	Tbilirubin	2.612	0.699	13.945	<0.001	13.621	3.459	53.638
	Dbilirubin	-0.659	1.775	0.138	0.711	0.517	0.016	16.790
	Sabit	-19.396	2.324	69.644	<0.001	0.000		
	3	Yaş	0.149	0.023	43.912	<0.001	1.161	1.111
Hemoglobin		-0.064	0.084	0.579	0.447	0.938	0.796	1.106
Ürik asit		0.383	0.131	8.529	0.003	1.467	1.134	1.897
LDL		0.064	0.009	51.826	<0.001	1.066	1.047	1.084
Tbilirubin		2.484	0.607	16.761	<0.001	11.985	3.650	39.360
Sabit		-19.435	2.318	70.321	<0.001	0.000		
4	Yaş	0.148	0.022	44.073	<0.001	1.159	1.110	1.211
	Ürik asit	0.371	0.130	8.156	0.004	1.450	1.124	1.870
	LDL	0.064	0.009	51.860	<0.001	1.066	1.047	1.084
	Tbilirubin	2.389	0.589	16.456	<0.001	10.905	3.438	34.589
	Sabit	-20.097	2.180	84.993	<0.001	0.000		

Tablo 2’de, geriye doğru eleme yönteminin 1. adımında modele bütün değişkenler dahil edilmiştir. HDL değişkeninin p-değeri daha önceden belirlenmiş olan kritere göre en büyük olarak saptanmış olup modele katkısının en az olduğu belirlenerek modelden çıkartılmış ve 2. adıma geçilmiştir. 2. adımda tekrar p değerlerinin karşılaştırılması sonucu direkt bilirubin değişkeninin diğer değişkenlere göre önemlilik seviyesi az bulunmuş ve elimine edilerek diğer adıma geçilmiştir. 4. adım sonucunda modelden atılacak değişken kalmamış olup yaş, ürik asit, LDL ve toplam bilirubin değişkenleri modele katkısı en yüksek olan değişkenler olarak belirlenmiştir (Tablo 2).

Tablo 3. Hosmer- Lemeshow Uyum iyiliği testi

Adım	Ki-kare	Serbestlik derecesi	p-değeri
1	9.175	8	0.328
2	15.134	8	0.057
3	14.949	8	0.060
4	15.913	8	0.054

Tablo 3’de elde edilen sonuçlara göre $p > 0.05$ olduğundan teorik modelin verileri iyi temsil ettiği elde belirlenmiştir. Bu test modelin bir bütün olarak uyumunu değerlendirmektedir.

Tablo 4. Model özeti

Adım	-2 Log olabilirlik	Cox & Snell R ²	Nagelkerke R ²
1	269.725	0.647	0.896
2	269.809	0.647	0.896
3	269.947	0.647	0.896
4	270.519	0.646	0.895

Tablo 4 incelendiğinde 1. adımda; bağımsız değişkenlerin, bağımlı değişkendeki toplam değişimin Cox-Snell'e göre %64.7 ve Nagelkerke'ye göre % 89.6'sını açıkladığı elde edilmiştir. Diğer adımlarda da benzer biçimde bağımsız değişkenlerin bağımlı değişken olan Koroner arter hastalığını etkileme dereceleri yer almaktadır.

Tablo 5. Sınıflandırma Tablosu

Adım	Gözlenen	Tahmin edilen			
		Koroner arter hastalığı		Sınıflandırma yüzdesi	
		Kontrol	Hasta		
1	Koroner arter hastalığı	Kontrol	142	28	83.5
		Hasta	34	136	80.0
	Toplam yüzde				81.8
2	Koroner arter hastalığı	Kontrol	142	28	83.5
		Hasta	34	136	80.0
	Toplam yüzde				81.8
3	Koroner arter hastalığı	Kontrol	142	28	83.5
		Hasta	34	136	80.0
	Toplam yüzde				81.8
4	Koroner arter hastalığı	Kontrol	142	28	83.5
		Hasta	26	144	84.7
	Toplam yüzde				84.1

Tablo 5'de 0.50 kritik değer (cut value) ve tahmin edilen olasılıklar yardımıyla sınıflandırma sonuçları yer almaktadır. Tablo 5'den görüldüğü gibi 4. adımda hasta olmayanların % 83.5'i, hasta olanların %84.7'si doğru tahmin edilmiştir. Genel olarak ise % 84.1'i doğru tahmin edilmiştir. Bu değerler lojistik regresyon analizi için oldukça uygun değerlerdir.

Tablo 6. Lojistik regresyon modeline dahil edilen değişkenler

Değişken	Katsayı	Standart hata	Wald istatistiği	Serbestlik derecesi	p-değeri	Odds oranı	% 95 güven aralığı	
							Alt sınır	Üst sınır
Yaş	0.148	0.022	44.073	1	<0.001	1.159	1.110	1.211
Ürik asit	0.371	0.130	8.156	1	0.004	1.450	1.124	1.870
LDL	0.064	0.009	51.860	1	<0.001	1.066	1.047	1.084
TBilirubin	2.389	0.589	16.456	1	<0.001	10.905	3.438	34.589
Sabit	-20.097	2.180	84.993	1	<0.001	0.000		

Tablo 6 incelendiğinde odds oranlarına göre; yaş arttıkça Koroner Arter Hastası olma oranının 1.159 kat, ürik asit değeri arttıkça 1.450 kat, düşük yoğunluklu lipoprotein (LDL) arttıkça 1.066 kat, toplam

bilirubin arttıkça ise 10.905 kat arttığı belirlenmiştir. Sonuç olarak yaş, ürik sit, LDL ve toplam bilirubin koroner arter hastalığını belirlemede önemli bağımsız değişkenler olarak bulunmuştur. Lojistik regresyon modeli ise aşağıdaki gibi elde edilmiştir.

$$P(KOAH = 1) = \left(1 + e^{-(-20.097 + 0.148 \text{ yaş} + 0.371 \text{ ürik asit} + 0.064 \text{ LDL} + 2.389 \text{ toplam bilirubin})}\right)^{-1}$$

Tablo 7. Jackknife yöntemine dayalı lojistik regresyon analizi sonuçları

Değişken	Orijinal lojistik regresyon analizi		Jackknife yöntemine dayalı lojistik regresyon analizi	
	Parameter tahmini	Standart hata	Parameter tahmini	Standart hata
Yaş	0.148	0.022	0.149	0.019
Ürik asit	0.371	0.130	0.391	0.112
LDL	0.064	0.009	0.063	0.002
Toplam Bilirubin	2.389	0.589	2.669	0.345
Sabit	-20.097	2.180	-19.065	1.997

Tablo 7’de Jackknife yeniden ornekleme sonucu kurulan model katsayıları ve geriye doğru eleme yöntemi sonucunda elde edilen model katsayıları verilmiştir. Jackknife yöntemine göre elde edilen model katsayılarının, logistic regresyon yöntemi ile elde edilen model katsayılarından çok düşük düzeyde sapmalar gösterdiği görülmüştür. Bunun yanı sıra, modellere ait standart hatalar incelendiğinde ise Jackknife yöntemine dayalı modelin standart hatalarının, geriye doğru eleme yöntemi sonucunda elde edilen lojistik regresyon modelinin standart hatalarından daha düşük olduğu belirlenmiştir.

5. Sonuç

Lojistik regresyon analizinde bağımsız değişkenler sürekli değişkenler ve kategorik olarak bir arada kullanılabilir. Bağımsız değişkenlerin dağılımları üzerinde hiçbir kısıtlayıcı varsayım bulunmamaktadır. Bu sebepten dolayı özellikle tıp alanında hastalığın birey üzerinde var olup olmadığının araştırıldığı çalışmalarda yoğun bir şekilde kullanılmaktadır. Bu amaç doğrultusunda yapılan bu çalışmada hastaneye başvuran bireylerde koroner arter hastalığının olup olmadığına ilişkin bir ayırtma modeli lojistik regresyon analizi kullanılarak elde edilmeye çalışılmıştır.

Lojistik Regresyon Analizinin kullanım amacı, en az değişkeni kullanarak en iyi uyuma sahip olacak şekilde bağımlı değişken ile bağımsız değişkenler arasındaki ilişkiyi tanımlayabilmek ve amaca yönelik kabul edilebilir bir model kurmaktır. Bu yöntemde, bağımsız değişkenlerin bağımlı değişken üzerindeki etkileri olasılık olarak hesaplanarak risk faktörlerinin olasılık olarak belirlenmesi sağlanır. Çalışmada Koroner arter hastası olma durumu en az değişkenle tahmin edilmeye çalışılmış ve bu değişkenlerin yaş, ldl ve trigliserit olduğu belirlenmiştir. Böylece az sayıda değişkenle zamandan kazanarak, işlem kolaylığı sağlanmıştır.

Bu aşamadan sonra yeniden ornekleme yöntemlerinden biri olan jackknife yöntemiyle örnekler elde edilmiş ve lojistik regresyon analizi sonucunda elde edilen modelle karşılaştırılıp hangisinin daha etkili parametre tahminleri elde ettiği ortaya konulmaya çalışılmıştır.

Çalışmada elde edilen sonuçlara göre yaş arttıkça Koroner Arter Hastası olma oranının 1.159 kat, ürik asit değeri arttıkça 1.450 kat, düşük yoğunluklu lipoprotein (LDL) arttıkça 1.066 kat, toplam bilirubin arttıkça ise 10.905 kat arttığı belirlenmiştir. Sonuç olarak yaş, ürik sit, LDL ve toplam bilirubin koroner arter hastalığını belirlemede önemli bağımsız değişkenler olarak bulunmuştur.

Orjinal lojistik regresyon analizi sonuçları ile jackknife yöntemine dayalı lojistik regresyon analizi sonuçları incelendiğinde ise katsayı tahminleri arasında çok az bir farklılık olduğu görülmüştür. Jackknife yönteminde parametre tahminlerinin standart hataları genellikle klasik ornekten elde edilen parametre tahminlerinin standart hatalarından, daha düşük bulunmuştur. Bu sonuç bize jackknife yöntemine dayalı parametre tahminlerinin oldukça etkin olduğu göstermektedir.

Delete d R-Algorithm

```
#This R code defines a function 'jack' for performing delete-d jackknife for logistic regression
#p is the no of cols in the data. p=8 then there is 1 dept. var and 7 indept vars
#d is the no of rows to be deleted
jack=function(dataset,p,d)
{
  n<- length(dataset[,1]) #the sample size
  u<- combn(n,d) #Assign the matrix of all possible combinations to u
  output<- matrix(0,ncol=p, nrow=ncol(u)) #define the output
  y<- dataset[,1] #the response vector
  x<- dataset[,2:p] #the matrix of covariates
  for (i in 1:(ncol(u)))
  {
    dd<-c(u[,i])
    yn<-y[-dd] #delete d rows of the independent var
    xn<-x[-dd,] #delete d rows of the dependent var
    logreg<- glm(yn~xn[,1]+xn[,2]+ xn[,3]+ xn[,4]+ xn[,5]+ xn[,6]+ xn[,7], family = binomial(link =
"logit"))
    coef<- logreg$coef
    output[i,]<- c(coef) #store the regression coefficients
  }
}
#This part can be used to obtain a jackknife estimate of the regression coefficients
u<- jack(dataset,8,1)
u[,1]
beta<-c(mean(u[,1]),mean(u[,2]),mean(u[,3]),mean(u[,4]),mean(u[,5]),mean(u[,6]),mean(u[,7]))
```

References

- [1] Gündoğdu, F., Özdemir, Ö., Sevimli, S., Açikel, M., Pirim, İ., Karakelleoğlu, Ş., et al. (2007). The relationship between interleukin-6 polymorphism and the extent of coronary artery disease in patients with acute coronary syndrome. *Arch. Turk. Soc. Cardiol.*, 35(5): 278–283.
- [2] Yin, Y., Li, J., Zhang, M., Wang, J., Li, B., Liu, Y., et al. (2013). Influence of interleukin-6 gene -174g>c polymorphism on development of atherosclerosis: a meta-analysis of 50 studies involving, 33.514 subjects. *Gene*, 529: 94–103.
- [3] Elsaid, A., Abdel-Aziz, A.F., Elmougy, R., Elwaseef, A.M. (2014). Association of polymorphisms G (-174) C in IL-6 gene and G (-1082) A in IL-10 gene with traditional cardiovascular risk factors in patients with coronary artery disease. *Indian J. Biochem. Biophys.*, 51: 282–292.
- [4] Alpar, R. (2011). *Uygulamalı Çok Değişkenli İstatistiksel Yöntemler*, Detay Yayıncılık, Ankara.
- [5] Kleinbaum, D., Klein, M. (2002). *Logistic Regression- A Self Learning Text*, II ed. New York, NY: Springer.
- [6] Berkson, J. (1944). Application of the logistic function to bio-assay, *J. Am. Stat. Assoc.*, 39(227): 357–365.
- [7] Coşkun, S., Kartal, M., Coşkun, A., Bircan, H. (2004). Lojistik regresyon analizinin incelenmesi ve dış hekimliğinde bir uygulaması, *Cumhuriyet Üniversitesi Dış Hekimliği Fakültesi Dergisi*, 7(1): 42–50.

- [8] Vupa, O., Çelikoğlu, C. (2006). Model building in logistic regression models about lung cancer data, *Anadolu Univ. J. Sci. Tech.*, 7(1): 127–141.
- [9] Hirashiki, A., Yamada, Y., Murase, Y., Hirashiki, A., Yamada, Y., Murase, Y. (2003). Association of gene polymorphisms with coronary artery disease in low- or high-risk subjects defined by conventional risk factors, *J. Am. Coll. Cardiol.*, 42(8): 1429–1437.
- [10] Horibe, H., Yamada, Y., Ichihara, S., Watarai, M., Yanase, M., Takemoto, K., et al. (2004). Genetic risk for restenosis after coronary balloon angioplasty, *Atherosclerosis*, 174(1): 181–187.
- [11] Çolak, C., Çolak, M.C., Orman, M.N. (2007). The Comparison of logistic regression model selection methods for the prediction of coronary artery disease, *The Anatol. J. Cardiol.*, 7(1): 6–12.
- [12] Quenouille, M. H. (1949). Approximate tests of correlation in time series. *J. Roy. Stat. Soc.*, 11, 18-44.
- [13] Tukey, J. W. (1958). Bias and confidence in not-quite large sample. *Ann Math Stat*, 29, 614.
- [14] Fenwick, I. (1979). Techniques in Market Measurement: The Jackknife. *Journal of Marketing Research*, 74 410-414.
- [15] Abdi, H. and Williams, J. L. (2010). Jackknife in Neil Salkind (Ed.). *Encyclopedia of Research Design*. <http://dx.doi.org/10.4135/9781412961288.n202>.
- [16] Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Boot-strap*. Chapman and Hall, New York.
- [17] Sahinler, S. and Topuz, D. (2007). Bootstrap and Jackknife Resampling Algorithm for Estimation of Regression Parameters. *J. Appl. Quant. Meth.* 2, 2:188-199.