İstatistik Araştırma Dergisi Cilt 04 No 03 Sayfa 52-62 Aralık 2005

REGRESYON ÇÖZÜMLEMESİNDE KAYIP VERİ SORUNU

Neslihan DEMİREL* Serdar KURT*

ÖZET

Kayıp veri çözümlemesinin konusu veri matrisindeki bazı değerlerin gözlenmemiş olmasıdır. Kayıp veri çözümlemesi özellikle uygulamalı istatistiğin çok önemli konularından birini oluşturmaktadır. Kayıp veriyi yok saymak, örneklemin rastgeleliğini bozarak yanlı parametre tahminleri elde edilmesine neden olabilmektedir. Regresyon analizi, tahmin amaçlı kullanılan önemli çok değişkenli istatistiksel analizlerin başında gelmektedir. Bu nedenle bu çalışmada, regresyon analizinde, bağımsız değişkenlerde kayıp veri mekanizması rassal kayıp (MAR) olacak şekilde, regresyon analizi varsayımlarının sağlandığı ve sağlanmadığı iki ayrı veri seti üzerinde benzetim çalışması yapılmıştır. Kayıp veri göz ardı edilebilir olduğunda model esaslı yöntemler arasında yer alan, EM algoritması ve çoklu atıf yöntemleri karşılaştırmalı olarak incelenmiştir. EM algoritmasının regresyon analizi varsayımlarının bozulmasından etkilenmediği, ancak çoklu atıf için, atıf sayısının arttırılması gerektiği sonucu elde edilmiştir.

Anahtar Kelimeler: *Çoklu Atıf EM Algoritması, Kayıp Veri, Regresyon Analizi.*

1. GİRİŞ

Çok değişkenli bir kitleden alınan örnekleme ait veri matrisinde gözlemler satırlarla, değişkenler sütunlarla temsil edilir. Kayıp veri çözümlemesinin konusu ise veri matrisindeki bazı değerlerin gözlenmemiş olmasıdır.

Kayıp veriyi yok saymak, örneklemin rastgeleliğini bozarak, yanlı tahminler elde edilmesine neden olabilmektedir. Kayıp veri çözümlemesi için geliştirilen yöntemlerden hangisinin uygun olduğuna karar verebilmek için, kayıp veri oluşumuna neden olan etkenlerin incelenerek, kayıp veri mekanizmasının bulunması gerekmektedir. Kayıp veri mekanizması, kayıp verilerin farklı nedenlerle ortaya çıkmasına bağlı olarak üç şekilde sınıflandırılır: tamamen rassal kayıp (missing completely at random – MCAR), rassal kayıp (missing at random – MAR) , rassal olmayan kayıp (not missing at random - NMAR).

* Dokuz Eylül Üniversitesi, Fen-Edebiyat Fakültesi, İstatistik Bölümü, Buca-İZMİR, TÜRKİYE
neslihan.ortabas@deu.edu.tr - serdar.kurt@deu.edu.tr

Y, nxK boyutlu tam veri matrisi olarak tanımlansın, $Y=(y_{ij})_{nxK}$. Burada $i=1,\ldots,n$ ile gözlem indisini, $j=1,\ldots,K$ ise değişken indisini temsil eder. Y'nin i. satırı $y_i=(y_{i1},\ldots,y_{iK})$ içinde yer alan y_{ij}, Y_j değişkeninin i. gözlemine ait değeri gösterir. R, nxK boyutlu kayıp veri gösterge matrisi olarak tanımlansın, $R=(r_{ij})_{nxK}$. y_{ij} kayıp ise $r_{ij}=1$, y_{ij} gözlenmiş ise $r_{ij}=0$ değerlerini alır. Kayıp veri mekanizması, Y verildiğinde R'nin koşullu dağılımı, $f(R|Y,\phi)$, olarak tanımlanır. Bu fonksiyonda ϕ bilinmeyen parametreleri temsil eder. Üç farklı şekilde sınıflandırılan kayıp veri mekanizması bu tanımlamalardan hareketle aşağıdaki gibi özetlenir.

MCAR: Bir gözlemin kayıp değeri, gözlenenler ve kayıp gözlemlerin değerlerine bağlı değilse, tüm Y ve ϕ'ler için $f(R|Y,\phi)=f(R|\phi)$'dır.

MAR: $Y_{gözlenen}$, Y'nin gözlenen bileşenlerini, $Y_{kayıp}$ ise kayıp bileşenlerini göstersin. Bir gözlemin kayıp değeri, gözlenenlere bağlı, kayıp gözlemlerin değerlerine bağlı değilse, tüm $Y_{kayıp}$ ve ϕ'ler için $f(R|Y,\phi)=f(R|Y_{gözlenen},\phi)$'dır.

NMAR: Bazı birimleri kayıp olan tek değişkenli rassal örneklem, bir veri seti olarak ele alınsın. $Y=(y_1,\ldots,y_n)'$ gösteriminde rassal değişkenin $i.$ birimin aldığı değer y_i'dir. Bu durumda $R=(R_1,\ldots,R_n)$ gösteriminde birimler gözlenmiş ise $R_i=0$, birimler gözlenmemiş ise $R_i=1$ değerini alır. (y_i,R_i)'nin bileşik dağılımının karşılıklı birimler için bağımsız olduğunu varsayalım, böylece özellikle gözlenen birimin olasılığı diğer birimler için Y'nin ya da R'nin değerlerine bağlı olmaz. O zaman $f(Y,R|\theta,\phi)=f(Y|\theta)f(R|Y,\phi)=\prod_{i=1}^{n} f(y_i|\theta)\prod_{i=1}^{n} f(R_i|y_i,\phi)$ elde edilir. Burada θ bilinmeyen parametreleri temsil eder ve y_i'nin yoğunluğu $f(y_i|\theta)$ ile gösterilir ve ikili gösterge R_i için Bernoulli dağılımının yoğunluğu $f(R_i|y_i,\phi)$'dir. y_i kayıp ise olasılık $Pr(R_i=1|y_i,\phi)$ ile gösterilir. Eğer kayıp değer Y'den bağımsız ise olasılık $Pr(R_i=1|y_i,\phi)=\phi$ bir sabite eşit, bu sabit y_i'den bağımsız olduğundan, kayıp veri mekanizması MCAR olur. Eğer y_i'nin bazı kayıp değerlerine bağlı çıkarsa NMAR olur. (Little ve Rubin, 2002).

Kayıp veri oluşumuna neden olan etkenler incelenerek bulunan kayıp veri mekanizmasına hangi çözümlemenin yapılacağına karar vermek için Afifi ve Elashoff (1966), Hartley ve Hocking (1971), Orchard ve Woodbury (1972), Dempster, Laird ve Rubin (1977), Little ve Rubin (1983a), Little ve Schenker (1994) ve Little (1997) gibi araştırmacılara ait çalışmalar incelendiğinde, kayıp veri çözümlemesi yöntemlerini 4 ana başlık altında sınıflamak mümkündür: Tamamen kayıtlanmış birim esaslı yöntemler (procedures based on completely recorded units), ağırlıklandırılmış yöntemler (weighting procedures), atıf esaslı yöntemler (imputation-based procedures) ve model esaslı yöntemler (model-based procedures) (Little ve Rubin, 2002).

Tamamen kayıtlanmış birim esaslı yöntemde bazı değişkenlerin bazı birimleri kayıtlanmamış ise kayıtlanmamış birimlere ait satırları tüm değişkenlerden kaldırarak, çözümlemeye geri kalan tam veri seti ile devam edilir. Çözümlemeden birim çıkarmak oldukça kolay bir yöntem olmasına rağmen bu durumda parametre kestirimleri önemli yanlılık gösterebileceğinden, yöntemin etkili olmadığı belirtilmektedir (Little ve Rubin, 2002).

Ağırlıklandırılmış yöntem daha yaygın olarak anket çalışmalarında cevaplamamadan kaynaklanan kayıp gözlemler için kullanılmaktadır. Örneklemin alt kümelerinde gözlenen değerlere ağırlıklar verilerek, kayıp birimler için her alt kümenin ağırlığına karşılık gelecek şekilde değerler atanır (Little ve Rubin, 2002).

Atıf esaslı yöntemde, kayıp gözlemlerin yeri doldurularak, tamamlanan veri seti üzerinde standart yöntemleri uygulama esasına dayanır. Hot deck atfı, regresyon atfı, ortalama atfı, vb. olmak üzere farklı atıf yöntemleri vardır. Uygun atıf yönteminin seçilmesinde dikkatli karar vermek gerekir (Little ve Rubin, 2002).

Model esaslı yöntem, kısmen kayıp veri seti için bilinmeyen parametrelerin tahmin edilmesinde en çok olabilirlik tahmin yöntemi gibi yöntemlerin kullanılması esasına dayanır. Bu yöntemin avantajı esnek olması, model varsayımları altında sonuçların sergilenip, değerlendirilmesidir (Little ve Rubin, 2002).

2. KAYIP VERİLER İÇİN GELİŞTİRİLEN ÇÖZÜM YÖNTEMLERİ

Kayıp veri mekanizması belirlendikten sonra, 1. bölümde bahsedilen 4 ana başlık altında toplanan, kayıp veri için geliştirilen çözüm tekniklerinden uygun olan kullanılır. Bu yöntemler, tam gözlemlerin kullanılması (listwise data deletion), gözlemlerin ya da değişkenlerin silinmesi (casewise data deletion), çiftler bazında veri silme (pairwise data deletion), yerine ortalamayı koyma (mean substution), regresyon atfı (regression imputation), hot deck atfı (hot deck imputation), beklenti maksimizasyonu algoritması-em algoritması (expectation maximization algorithm) ve çoklu atıf (multiple imputation) olarak sıralanabilir.

Bu çalışmada ise kayıp veri mekanizması rassal kayıp (MAR) olan veri seti üzerinde çalışılmıştır. Kayıp veri göz ardı edilebilir olduğunda model esaslı yöntemler arasında yer alan ve bu çalışmada kullanılan EM algoritması ve çoklu atıf yöntemlerine kısaca değinilmiştir.

2.1 EM Algoritması

EM algoritması özellikle verinin bir kısmı kayıp olduğunda tahmin amaçlı kullanılan genel iteratif bir algoritmadır. Her bir iterasyon iki adımdan oluşur. Beklenti ve maksimizasyon. Bu adımlar aşağıdaki gibi tekrarlanır.

1. Kayıp değerler, tahmin değerleri ile yenilenir.
2. Parametreler tahmin edilir.
3. Yeni parametre tahminlerinin doğru olduğu varsayımı altında, kayıp değerler yeniden tahmin edilir.
4. Parametreler yeniden tahmin edilir

Bu iterasyon, ardışık iterasyondan elde edilen tahminlerin arasındaki fark önemsenmeyecek biçimde azalana kadar devam eder. Ortalama, standart sapma, korelasyon, kovaryans matrisi gibi özet istatistikler, standart doğrusal model yazılımlarından elde edilir (Schafer, 1997)

Çok değişkenli normal dağılan ve göz ardı edilebilen kayıp veri mekanizmasına sahip, kayıp gözlemleri bulunan bir örneklemin ortalama ve kovaryans matrisi en çok olabilirlik tahminleri ile elde edilir.

$Y=(Y_1,Y_2\ldots,Y_K)$, ortalaması $\mu=(\mu_1,\mu_2,\ldots,\mu_K)$ ve kovaryans matrisi $\sum=(\sigma_{jk})$ olan K-değişkenli normal dağılsın. $Y=(Y_{gözlenen},Y_{kayıp})$ yazıldığında gözlenen değerler seti, $Y_{gözlenen}$ ile, kayıp değerler seti $Y_{kayıp}$ ile gösterilir. $Y_{gözlenen}=(y_{gözlenen,1},y_{gözlenen,2},\ldots,y_{gözlenen,n})$ gösteriminde $y_{gözlenen,i}$: i. gözlem için gözlenen değişkenler setini temsil eder. $(i=1,\ldots,n)$

Gözlenen değerler için olabilirlik fonksiyonu;

$$l(\mu,\Sigma\,|\,Y_{gözlenen})=sabit-\frac{1}{2}\sum_{i=1}^{n}ln\left|\Sigma_{gözlenen,i}\right|-\frac{1}{2}\sum_{i=1}^{n}(y_{gözlenen,i}-\mu_{gözlenen,i})'\Sigma^{-1}_{gözlenen,i}(y_{gözlenen,i}-\mu_{gözlenen,i}) \quad (1)$$

burada $\mu_{gözlenen,i}$ ve $\Sigma_{gözlenen,i}$ Y'nin gözlenen bileşenlerinin i. gözlemi için ortalama ve kovaryans matrislerini temsil eder.

(1) eşitliğini EM algoritmasında en büyüklemek üzere, Y'nin kayıp gözlem içermediği varsayımı altında, Y üssel ailesi için Y_j ve Y_k $(j,k=1,\ldots k)$ rassal değişkenlerine ait yeterli istatistikleri

$$S=\left(\sum_{i=1}^{n}y_{ij};\quad j=1,\ldots,K;\quad \sum_{i=1}^{n}y_{ij}y_{ik};\quad j,k=1,\ldots,K\right) \quad (2)$$

dir. t. iterasyonda EM, parametre tahmini için $\theta^{(t)}=(\mu^{(t)},\Sigma^{(t)})$ ile gösterilir. Algoritma E adımda

$$E(\sum_{i=1}^{n} y_{ij} \mid Y_{gözlenen}, \theta^{(t)}) = \sum_{i=1}^{n} y_{ij}^{(t)}, \quad j = 1,...,K$$

ulaşır ve

$$E(\sum_{i=1}^{n} y_{ij} y_{ik} \mid Y_{gözlenen}, \theta^{(t)}) = \sum_{i=1}^{n} (y_{ij}^{(t)} y_{ik}^{(t)} + c_{jki}^{(t)}); \quad j,k = 1,...,K$$

olur. Burada

$$y_{ij}^{(t)} = \begin{cases} y_{ij} & , y_{ij} \text{ gözlenmis ise;} \\ E(y_{ij} \mid y_{gözlenen,i}\theta^{(t)}) & , y_{ij} \text{ kayıp ise;} \end{cases} \tag{3}$$

ve

$$c_{jki}^{(t)} = \begin{cases} 0 & , y_{ij} \textit{ veya } y_{ik} \text{ gözlenmis ise;} \\ \mathrm{cov}(y_{ij}, y_{ik} \mid y_{gözlenen,i}\theta^{(t)}) & , y_{ij} \textit{ ve } y_{ik} \text{ kayıp ise;} \end{cases} \tag{4}$$

olarak gösterilir.

Kayıp y_{ij} değerleri, $y_{gözlenen,i}$ değerlerinin bilinmesi durumunda y_{ij}'nin koşullu ortalaması ile yer değiştirir.

EM algoritmasının M adımında parametrelerin yeni tahminleri $\theta^{(t+1)}$ tahmin edilir. (Little ve Rubin, 2002). Bunlar,

$$\mu_j^{(t+1)} = n^{-1} \sum_{i=1}^{n} y_{ij}^{(t)}, \quad j = 1,..., K; \tag{5}$$

$$\sigma_{jk}^{(t+1)} = n^{-1} E(\sum y_{ij} y_{ik} \mid Y_{gözlenen}) - \mu_j^{(t+1)} \mu_k^{(t+1)}$$

$$= n^{-1} \sum_{i=1}^{n} [(y_{ij}^{(t)} - \mu_j^{(t+1)})(y_{ik}^{(t)} - \mu_k^{(t+1)}) + c_{jki}^{(t)}]; \quad j,k = 1,...,K$$

ile elde edilir.

2.2 Çoklu Atıf

Her kayıp değer yerine tek bir değer atanması yerine, çoklu atıfta her kayıp değere birkaç kez atama yapılarak, birkaç tam veri seti elde edilir. Analizler ayrı ayrı yapılırak, sonuçlar birleştirilir. Rubin'in notasyonuna göre, gözlenen değerler $Y_{gözlenen}$, kayıp değerler $Y_{\text{kayıp}}$ ile gösterilir. Kitle niceliği Q'nun sonsal yoğunluğu aşağıdaki gibi yazılır.

$$h(Q \mid Y_{gözlenen}) = \int g(Q \mid Y_{gözlenen}, Y_{\text{kayıp}}) F(Y_{\text{kayıp}} \mid Y_{gözlenen}) dY_{\text{kayıp}} \tag{6}$$

Burada $f(.)$ kayıp değerlerin sonsal yoğunluk fonksiyonu ve $g(.)$ θ'nın tam veri seti için sonsal yoğunluk fonksiyonudur. Çoklu atıflar, kayıp verinin sonsal dağılımlarından benzetim ile seçilir.

Tam veri istatistikleri olan $\hat{Q}$ ile U, s tam veriden elde edilen $\hat{Q}_{*1},...,\hat{Q}_{*s}$ ile $U_{*1},...,U_{*s}$ hesaplanır. Tekrarlı atıf tahminleri

$$\overline{Q}_s = \sum_{l=1}^{s} Q_{*l}/s \tag{7}$$

ve $\overline{Q}_s$ 'in varyans-kovaryansı

$$T_s = \overline{U}_s + \frac{s+1}{s} B_s \tag{8}$$

olur. Burada

$$\overline{U}_s = \sum_{l=1}^{s} U_{*l}/s \tag{9}$$

atıf-içi değişkenlik ve

$$B_s = \sum_{l=1}^{s} (Q_{*l} - \overline{Q}_s)(Q_{*l} - \overline{Q}_s)'/(s-1) \tag{10}$$

atıflar-arası değişkenlik olarak isimlendirilir.

s'nin büyük değerleri için tekrarlı-atıf çıkarımlarından $(Q - \overline{Q}_s)$ normal dağılır, varyans-kovaryans matrisi T_s ile gösterilir. $s = \infty$ olduğunda, $(Q - \overline{Q}_s)$'nin dağılımı

$$(Q - \overline{Q}_\infty) \sim N(0, T_\infty) \tag{11}$$

olur. Burada $T_\infty = \overline{U}_\infty + B_\infty$ dur. (Atkinson ve Cheng, 2000).

Çoklu atıf ve EM algoritması karşılaştırılmalı olarak özetlenirse, EM algoritmasında E adımında kayıp değerler atanır ve tamamlanmış veriyle M adımda parametreler tahmin edilir. Böylece, kayıp değerler elde edilerek parametrelerin en çok olabilirlik tahminleri bulunur. En çok olabilirliğin kayıp veri yaklaşımında çok önemli bir yeri olmasına rağmen, doğrusal ve logaritmik doğrusal modellerde bazı kısıtları vardır. Çoklu atıf ise en çok olabilirlik ile aynı özelliklere sahiptir. Aynı zamanda her tür veriye ve her tür modele uygulanabilir. Tek dezavantajı çoklu atıfın her kullanımında farklı tahminler elde edilebilir olmasıdır (McLachlan ve Krishnan, (1997)). Farklı araştırmacılar aynı veri için aynı yöntemle farklı değerler elde edebilir, istenen aradaki farkın önemsenmeyecek kadar az olmasıdır (Allison, 2002).

3. BENZETİM ÇALIŞMASI

Atkinson ve Cheng (2000) çalışmasında EM algoritması ve çoklu atıf yöntemlerini karşılaştırmak üzere bir benzetim çalışması uygulamıştır. Bu çalışmada 4 boyutlu çok değişkenli normal dağılan bir veri setinden 100 ve 200 birimlik örneklemler çekerek %10, %20, %30 ve %40 oranında rastgele kayıp gözlemler yaratmışlar ve yöntemleri karşılaştırmışlardır. Bu çalışmamızda, Atkinson ve Cheng'in (2000) çalışmasına ek olarak regresyon analizi varsayımından sapma olması durumunda yöntemlerin nasıl sonuçlar vereceği vurgulanmak istenmiştir. Çalışmada, ilk olarak Minitab paket programında simetrik ve çarpık olmak üzere iki ayrı kitle yaratıldı. Simetrik kitle için, üç bağımsız değişkenli $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$ regresyon modeli kuruldu. Bağımsız değişkenlerden oluşan X matrisi çoklu normal dağılımdan $MN(0, I_4)$ türetildi. Regresyon modelinin tüm parametrelerine 1 değeri verildi. Hata terimi 0 ortalamalı 1 varyanslı normal dağıldı. ($\varepsilon_i \sim N(0,1)$). İkinci uygulama için (çarpık kitle için) yine üç bağımsız değişkenli bir regresyon modeli kuruldu. Çarpık bir kitle yaratabilmek için X matrisi 2 serbestlik dereceli χ^2 dağılımdan türetildi. Yine regresyon modelinin tüm parametrelerine 1 değeri verildi. Bu kitlelerden 100 birimlik genişlikte örneklemler çekildi. Bu örneklemlerde her bir bağımsız değişken üzerinde eşit oranda kayıp sağlanması amacıyla tüm X matrisi üzerinde %12, %24 ve %36'lık kayıp değer oluşturuldu. Kayıp değerli örneklemlere EM algoritması ve çoklu atıf yöntemleri uygulandı. Çoklu atıflar arasında da karşılaştırma yapabilmek amacıyla 1'den 10'a kadar olan atıf sayıları arasından 2, 5 ve 10 atıf sayıları seçildi. Regresyon katsayıları ve modelin standart hatası ($\sqrt{Hata\ Kareler\ Ortalaması}$) karşılaştırma kriteri olarak kullanılmak üzere hesaplandı. Bu amaçla yapılan benzetim çalışmasının adımları aşağıdaki gibidir.

1. Simetrik bir kitle türetildi.
2. n=100 birimlik örneklem çekildi.
3. X matrisinde %12'si kayıp değer olarak atandı.
4. 2'li , 5'li ve 10'lu Çoklu Atıf ve EM algoritması uygulandı.
5. Regresyon katsayıları ve modelin standart hatası hesaplandı.
6. 2. adıma dönerek bu süreç 300 kez tekrarlandı.

Çalışma, 3. adımda %24 ve %36 kayıp değer olacak şekilde tekrar uygulandı. Aynı işlemlere 1. adımda türetilen kitle çarpık kitle olacak şekilde, tekrar devam edildi.

Minitab paket programında kitlenin türetilmesi, makro program aracılığı ile eksik veri türetilmesinde yararlanıldı. Bu örneklemler SOLAS paket programına aktarılarak çoklu atıflar, SPSS paket programına aktarılarak ise EM algoritması uygulandı. Uygulanan yöntemler sonunda kayıp değerlerin yerine tahmin değerleri konularak Minitab paket programında makro program aracılığıyla regresyon çözümlemesi yapılarak, regresyon katsayıları ve modelin standart hatası kayıt edildi. Yapılan çalışmadan elde edilen sonuçlar aşağıdaki tablolarda sunulmuştur.

Tablo 1. Simetrik Veri için 300 Tekrardan Sonra Elde Edilen Parametre Tahminlerinin Ortalaması, Standart Sapma Değerleri ile Modelin Standart Sapması.

Kayıp veri oranı %	Yöntem	$E(\hat{\beta}_0)$ $(S_{\hat{\beta}_0})$	$E(\hat{\beta}_1)$ $(S_{\hat{\beta}_1})$	$E(\hat{\beta}_2)$ $(S_{\hat{\beta}_2})$	$E(\hat{\beta}_3)$ $(S_{\hat{\beta}_3})$	Standart Hata
12	ÇA(2)	1.00234 (0.11360)	0.89951 (0.13070)	0.89484 (0.13485)	0.89360 (0.11900)	0.2679
	ÇA(5)	1.00614 (0.12378)	0.90321 (0.14728)	0.88632 (0.14077)	0.88617 (0.14478)	0.2636
	ÇA(10)	1.00861 (0.11324)	0.88136 (0.14065)	0.88074 (0.14200)	0.89486 (0.13295)	0.3066
	EM	0.99596 (0.09739)	0.97504 (0.11040)	0.97468 (0.10519)	0.97649 (0.10353)	0.2036
24	ÇA(2)	1.01883 (0.13665)	0.77319 (0.16179)	0.75485 (0.15405)	0.75536 (0.17101)	0.3489
	ÇA(5)	1.01902 (0.13837)	0.77400 (0.16151)	0.74761 (0.16345)	0.78096 (0.15766)	0.3433
	ÇA(10)	1.01577 (0.14104)	0.75798 (0.15866)	0.74450 (0.17125)	0.76661 (0.16511)	0.3591
	EM	1.01670 (0.12940)	0.76839 (0.25075)	0.75050 (0.26198)	0.76876 (0.25945)	0.6803
36	ÇA(2)	1.02200 (0.14122)	0.65712 (0.17048)	0.64930 (0.16580)	0.66161 (0.17325)	0.3673
	ÇA(5)	1.02815 (0.14659)	0.65937 (0.16465)	0.64781 (0.16662)	0.65456 (0.18045)	0.4222
	ÇA(10)	1.03076 (0.14862)	0.65609 (0.17962)	0.64791 (0.18280)	0.64215 (0.17473)	0.4154
	EM	1.01221 (0.13198)	0.76841 (0.14844)	0.76253 (0.14292)	0.77451 (0.15264)	0.3391

Kitle parametrelerinin değeri 1 olduğundan, Tablo 1 incelendiğinde $\hat{\beta}_i$ katsayılarının ortalamaları, 1'e en yakın değerleri ve modelin en küçük standart hatasını %12 ve %36'lık kayıplarda, EM algoritmasından elde etmiştir. %24'lük kayıpta ise 5'li çoklu atıf modelin standart hatasını en küçük vermiştir. Atkinson ve Cheng (2000) çalışmasında $\hat{\beta}_i$ katsayılarının ortalamalarının 1'e en yakın değerlerini çoklu atıfta bulmuştur. 5'li ve 10'lu çoklu atıfların, 2'li çoklu atıfa göre daha iyi sonuçlar verdiğini gözlemlemiştir. $\hat{\beta}_0$ değerlerinin ortalaması için 1'den büyük değerler, $\hat{\beta}_1, \hat{\beta}_2$ ve $\hat{\beta}_3$'te 1'den düşük değerler elde edilmiştir. Çoklu atıflar kendi içinde değerlendirilirse aralarında çok önemli bir fark gözlenmemekle beraber, genelde 5'li çoklu atıf kitle değerine daha yakın sonuç vermiştir.

Tablo 2. Çarpık Veri için 300 Tekrardan Sonra Elde Edilen Parametre Tahminlerinin Ortalaması, Standart Sapma Değerleri ile Modelin Standart Sapması.

Kayıp veri oranı %	Yöntem	$E(\hat{\beta}_0)$ $(S_{\hat{\beta}_0})$	$E(\hat{\beta}_1)$ $(S_{\hat{\beta}_1})$	$E(\hat{\beta}_2)$ $(S_{\hat{\beta}_2})$	$E(\hat{\beta}_3)$ $(S_{\hat{\beta}_3})$	Standart Hata
12	ÇA(2)	0.98865 (0.11860)	0.92644 (0.14967)	0.88079 (0.14408)	0.85505 (0.14429)	0.3788
	ÇA(5)	0.98816 (0.11946)	0.92963 (0.15634)	0.87729 (0.14150)	0.84995 (0.14253)	0.3843
	ÇA(10)	0.98227 (0.11797)	0.92023 (0.15610)	0.88627 (0.13554)	0.85671 (0.14329)	0.3734
	EM	0.99248 (0.11086)	0.96604 (0.13602)	0.90449 (0.13278)	0.88561 (0.12704)	0.3499
24	ÇA(2)	0.99684 (0.14136)	0.79486 (0.16362)	0.76363 (0.15342)	0.75197 (0.16239)	0.4417
	ÇA(5)	0.99279 (0.13227)	0.80806 (0.17032)	0.75760 (0.16906)	0.75450 (0.18137)	0.4479
	ÇA(10)	1.00350 (0.12876)	0.81109 (0.18312)	0.76535 (0.16704)	0.74960 (0.16559)	0.4391
	EM	0.995387 (0.14343)	0.77138 (0.34559)	0.73023 (0.32510)	0.729011 (0.30076)	0.8214
36	ÇA(2)	0.98886 (0.14097)	0.67761 (0.19223)	0.66218 (0.19182)	0.64490 (0.18511)	0.5248
	ÇA(5)	0.98766 (0.15672)	0.68323 (0.20389)	0.66941 (0.18420)	0.64564 (0.19014)	0.4848
	ÇA(10)	0.99866 (0.14883)	0.66894 (0.20154)	0.65555 (0.20469)	0.63507 (0.19116)	0.5589
	EM	0.98744 (0.14389)	0.80658 (0.15471)	0.78319 (0.14598)	0.76185 (0.15699)	0.4572

Kitle parametrelerinin değeri 1 olduğundan, Tablo 2 incelendiğinde $\hat{\beta}_i$ katsayılarının ortalamaları, 1'e en yakın değerleri değerleri ve modelin en küçük standart hatasını %12 ve %36'lık kayıplarda, EM algoritmasından elde etmiştir. %24'lük kayıpta ise 10'lu çoklu atıf modelin standart hatasını en küçük vermiştir. Çarpık kitlede tüm $\hat{\beta}_i$ değerlerinin ortalaması 1'den küçük değerler almıştır. Çoklu atıflar kendi içinde değerlendirilirse aralarında çok önemli bir fark gözlenmemekle beraber, genelde 10'lu çoklu atıf kitle değerine daha yakın sonuç vermiştir.

4. SONUÇ

Bir veri setinde bulunan kayıp veriyi yok saymak, örneklemin rastgeleliğini bozarak, yanlı tahminler elde edilmesine neden olabilmektedir. Bu nedenle, bu çalışmada, regresyon analizinde, bağımsız değişkenlerde ortaya çıkan rassal kayıplar EM algoritması ve çoklu atıf yöntemi ile giderilmeye çalışılmıştır. Bu amaçla yapılan benzetim çalışmasında Atkinson ve Cheng'in (2000) çalışması temel alınmış ve ek olarak, regresyon analizinde hataların normal dağılması gerektiği varsayımından sapma olması durumunda, yöntemlerin nasıl sonuçlar vereceği vurgulanmak istenmiştir.

Yöntemleri karşılaştırmak için, regresyon katsayılarının beklenen değerinin 1'e yakın, regresyon katsayılarının standart hatalarının ve modelin standart hatasının küçük çıkması kriterleri kullanıldığında simetrik kitle için %12 ve %36'lık kayıpta EM algoritmasının en iyi sonucu verdiği gözlenmiştir. %24'lük kayıpta ise 5'li çoklu atıf yöntemi modelin hatasını en küçük yapabilmiştir. Çarpık kitle için değerlendirme yapıldığında yine %12 ve %36'lık kayıpta EM algoritması en iyi sonucu verirken, %24'lük kayıpta ise 10'lu Çoklu Atıf modelin hatasını en küçük yapabilmektedir.

Sonuç olarak regresyon analizinde hataların normal dağılması gerektiği varsayımından sapma olması durumunda, EM algoritması bu durumdan etkilenmemektedir. Ancak çoklu atıf için atıf sayısının arttırılması gerektiği söylenebilir.

KAYNAKLAR

AFIFI, A.A. and ELASHOFF, R.M. (1966). *Missing Observations in Multivariate Statistics: Review of the Literature,* J. Am. Statist. Assoıc. 61, 595-604.

ALLISON, P.D. (2002). *Missing Data, Sage Publications,* USA

ATKINSON, A.C. and CHENG, T-C. (2000). *On Robust Linear Regression with Incomplete Data,* Computational Statistics & Data Analysis, 33, 361.

DEMPSTER, A.P., LAIRD, N.M. and RUBIN, D.B. (1977). *Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion),* J. Roy. Statist. Soc. B39, 1-38.

HARTLEY, H.O. and HOCKING, R.R. (1971). *The Analysis of Incomplete Data* , Biometrics 14, 174-194.

LITTLE, R. J. A. (1997). *Biostatistical Analysis with Missing Data,* in Encyclopedia of Biostatistics (P. Armitage and T. Colton, eds.) London: Wiley.

LITTLE, R. J. A. and RUBIN, D. B. (1983a). *IncompleteData,*in Encylopedia of Biostatistics *(P. Armitage and T. Colton, eds.) London: Wiley.*

LITTLE, R. J. A. and RUBIN, D. B. (2002), 2nd Ed., *Statistical Analysis with Missing Data,* A John Wiley & Sons, Inc. USA.

LITTLE, R.J.A. and SCHENKER, N. (1994). *Missing Data in Handbook for Statistical Modeling in the Social and Behavioral Sciences (G. Arminger, C.C. Clogg, and M.E. Sobel, eds.)* pp.39-75. NewYork: Plenum.

MCLACHLAN, G.J. and KRISHNAN, T. *(1997. The EM Algorithm and Extensions. New York: Wiley.*

ORCHARD, T. and WOODBURY, M.A. *(1972). A Missing Information Principle: Theory and Applications,* Proceedings of the 6th Berkeley Symposium on Mathematics, Statistics, and Probability, Volume 1, 697-715.

RUBIN, D.B. *(1976). Inference and Missing Data. Biometrika 63, 581-592.*

SCHAFER, J.L. *(1997). Analysis of Incomplete Multivariate Data,* Chapman & Hall, USA.

THE PROBLEM OF MISSING DATA IN REGRESSION ANALYSIS

ABSTRACT

The subject of missing data analysis consists of a data matrix in which some of the values in the matrix are not observed. Missing data analysis is one of the most important topics in applied statistics. It destroys the randomness of the sample and causes serious bias in the parameter estimate. The regression analysis is one of the most important procedures used for estimation in multivariate statistical analysis. For this reason, in this study, missing data mechanism designed by missing at random (MAR) for independent variables in regression analysis in two different data sets; one that verifies, one that violates regression assumptions; is used. When missing data can be ignored, model based methods that EM algorithm and multiple imputation method are compared. EM algorithm is not affected by the violation of regression assumptions but for multiple imputation number of imputations needs to be increased.

Key Words: *EM Algorithm, Missing Data, Multiple Imputation, Regression Analysis.*