İstatistikçiler Dergisi: İstatistik & Aktüerya

Journal of Statisticians: Statistics and Actuarial Sciences

IDIA 9, 2016, 2, 87-97

Geliş/Received:21.10.2016, Kabul/Accepted: 25.12.2016

www.istatistikciler.org

Araştırma Makalesi / Research Article

Sağlık Sigortasında Toplam Hasar Tutarının Kestirimi için Tek-kısım ve İki-kısım Modellerin Karşılaştırılması

Aslıhan Şentürk Acar

Hacettepe Üniversitesi, Aktüerya Bilimleri Bölümü, 06800, Beytepe, Ankara, Türkiye
aslihans@hacettepe.edu.tr

Uğur Karabey

Hacettepe Üniversitesi, Aktüerya Bilimleri Bölümü, 06800, Beytepe, Ankara, Türkiye
ukarabey@hacettepe.edu.tr

Öz

Doğrusal regresyon ve lognormal modelleri sağlık harcamalarının analizinde geleneksel olarak kullanılan tek-kısım modellerdir. İki-kısım modeller, tek-kısım modellere alternatif olarak sıklıkla tercih edilmektedir. İki-kısım modellerin ilk kısmında bireyin sağlık hizmetinden yararlanma olasılığı; ikinci kısmında sağlık harcamaları modellenmektedir. Bu çalışmada, bir hesap döneminde sağlık sigortası hasarlarında gözlenen yanıtların yalnızca toplam hasar tutarları olduğu durum göz önüne alınarak bireylerin toplam hasar tutarının tek-kısım modeller ve iki-kısım modeller ile kestirimine odaklanılmıştır. Bu amaçla, Türkiye'de faaliyet gösteren özel bir sigorta şirketinden alınan sağlık sigortası verisi kullanılarak, hata kareler ortalamasının karekökü ve ortalama mutlak hata kriterlerine göre aday modellerin kestirim performansı karşılaştırılmış, sonuçlar tartışılmıştır.

Anahtar sözcükler: Tek-kısım model, iki-kısım model, genelleştirilmiş doğrusal model, sağlık sigortası.

Abstract

Comparing One-Part Models and Two-Parts Models for the Prediction of Total Claim Amount in Health Insurance

Linear regression and lognormal models are the one-part models that are traditionally used to analyse health care expenditures. The method of two-part models is frequently used as an alternative to one-part models. Probability of health care utilization is modeled in the first part and the health care expenditures are modeled in the second part of two-part models. Considering the aggregate losses as the only observed responses of health insurance claims in an accounting period, we focus on the prediction of total claim amount of individuals using one-part models and two-part models. Accordingly, by using a private health insurance data set from a Turkish insurance company, predictive performance of candidate models are compared in terms of root mean square error (RMSE) and mean absolute error (MAE) criteria and the results are discussed.

Keywords: *One-part model, two-part model, generalized linear model, health insurance.*

1. Giriş

Artan sağlık harcamaları ile birlikte sağlık hizmetlerinde risk değerlendirme tekniklerinin kullanılması ve geliştirilmesi ülkeler için gittikçe önem arz etmektedir. Bireylerin demografik, ekonomik ve sağlık koşulları gibi bilgileri kullanılarak sağlık harcamalarının kestiriminin yapılması risk değerlendirme planlarının en önemli kısmıdır [3]. Sağlık harcamaları verisinde iki temel özellik gözlenmektedir. Bunlardan biri, gözlem dönemi boyunca sağlık hizmetinden yararlanmayan bireyler nedeniyle oluşan yüksek sayıda sıfır gözlemler diğeri ise pozitif harcamaların sağa çarpık dağılımıdır. Sağlık harcamalarının kestiriminde geleneksel olarak uygulanan doğrusal regresyon yönteminde verinin karma dağılım yapısı göz ardı edilmekte ve parametre tahmininde kullanılan en küçük kareler (EKK) yönteminden yanlı ve etkin olmayan tahminler elde edilmektedir [5]. Geleneksel olarak kullanılan diğer bir yöntem, harcamaların sağa çarpık dağılımının daha simetrik hale getirilmesi için veriye logaritmik dönüşüm uygulandığı ve EKK yöntemi ile parametre tahminlerinin elde edildiği lognormal modeldir.

EKK yöntemine alternatif olarak iki-kısım model, yatay-kesitsel sağlık hizmeti kullanımı ve sağlık harcamaları verisinin modellenmesinde sıklıkla tercih edilen bir yöntemdir. İki-kısım model adından da anlaşılacağı gibi iki ayrı kısımdan oluşmaktadır. İlk kısımda, sağlık hizmetinden yararlanma olasılığı modellenmektedir. Bu kısımda ikili rastlantı değişkeninin modellenmesi için genellikle probit ve lojistik regresyon kullanılmaktadır. İkinci kısımda sağlık hizmeti kullanımı koşulunda pozitif harcamalar modellenmektedir. Bu kısım için doğrusal regresyon, lognormal model ve genelleştirilmiş doğrusal model (GDM) gibi yaklaşımlar kullanılmaktadır [3;7;17].

Hayat dışı sigorta poliçesi, sigortacının bir yıl gibi belirli bir zaman döneminde poliçe sahibinin önceden bilinemeyen zararlarını tazmin etmesi karşılığında poliçe sahibinden prim aldığı, sigorta şirketi ile poliçe sahibi arasında yapılan bir anlaşmadır. Sağlık harcamaları verisi ile hayat-dışı sigorta hasar verisinin ortak özelliği yüksek sayıda sıfır değerleri ve pozitif harcamaların/hasar tutarlarının sağa çarpık dağılım yapısıdır. Kasko sigortası ve sağlık sigortası gibi branşlarda portföyde gözlenen sıfır değerlerinin bir nedeni poliçe dönemi boyunca hasar yapmayan bireyler olurken, sigorta şirketinin uyguladığı muafiyet ve poliçe sahiplerinin hasarsızlık indiriminden faydalanmak için küçük meblağlardaki hasarları şirkete bildirmemesi de diğer nedenlerdendir.

Hayat-dışı sigortada hasar modellemesinin bir çıktısı olan hasarların kestirimci dağılımı aktüeryal karar verme sürecinin temelini oluşturmaktadır. Kestirimci modeller risk sınıflandırması ve prim belirlenmesinde kullanılmaktadır [18]. Sigorta şirketlerinde bir yıl gibi belirli bir muhasebe döneminde farklı biçimlerde hasar verisi kaydedilmektedir. Kaydedilen hasar verisi yapısına göre farklı modelleme ve fiyatlama teknikleri kullanılmaktadır. Hasar verisi aşağıdaki biçimlerde olabilir:

1. Sadece toplam hasar tutarı bilgisi olabilir.
2. Hasar sayısı ve toplam hasar bilgisi olabilir.
3. Her bir hasar hakkında bilgi sağlayan, bireysel hasar tutarları ve hasar sayısından oluşan detaylı veri olabilir.

Birinci tip verinin modellenmesinde tek-kısım veya iki-kısım modellerden yararlanılabilir. İkinci ve üçüncü tip veride hasar frekansı veya hasar sayısı için bir dağılım seçilir. Bu dağılım genellikle Poisson veya negatif binom dağılımı olmakta ve sayı değişkeninin modellenmesinde sıklıkla genelleştirilmiş doğrusal modeller kullanılmaktadır [4]. İkinci tip veride toplam hasar tutarının hasar sayısına bölünmesiyle elde edilen ortalama hasar tutarları (hasar şiddeti) modellenir. Hasar tutarlarının sağa çarpık dağılımı nedeniyle bu değişkenin modellenmesinde sıklıkla kullanılan yöntemlerden biri gamma GDM'dir [11]. Üçüncü tip veride bireysel hasar tutarları, hasar sayısı koşulunda lognormal model ve GDM'ler ile modellenebilir [9]. Alternatif olarak her bir hasarda oluşan tutarların modellenmesi için doğrusal karma model [8] veya genelleştirilmiş doğrusal karma modellerden [1] yararlanılabilir. Hasar tutarlarının modellenmesinde sıklıkla kullanılan dağılımlar gamma ve lognormal dağılımdır.

Bu çalışmada, herhangi bir hayat-dışı sigorta branşında hasar bilgisi olarak bireylerin yalnızca toplam hasar tutarı bilgisinin bulunduğu durum ele alınmıştır. Poliçe sahiplerinin poliçe dönemi başında bilinen

bilgileri kullanılarak yıl boyunca oluşabilecek toplam hasar tutarının kestirimi için tek-kısım modeller ve iki-kısım modellerin kullanılması ve modellerin kestirim performanslarına göre karşılaştırılması amaçlanmıştır. Bu amaçla Türkiye'de faaliyet gösteren özel bir sigorta şirketinden alınan sağlık sigortası verisi kullanılarak tek-kısım ve iki-kısım modeller kestirim performanslarına göre karşılaştırılmıştır. Modellerin kestirim performansının ölçülmesinde iki kriter kullanılmıştır: Hata kareler ortalamasının karekökü ve ortalama mutlak hata.

Çalışmanın ikinci kısmında kullanılan modeller açıklanmıştır. Üçüncü kısımda uygulama çalışması yapılmış, kullanılan veri açıklanmış, aday modeller veriye uygulanmış ve model doğrulama yöntemi kullanılarak modeller kestirim performanslarına göre karşılaştırılmıştır. Dördüncü kısımda sonuçlar tartışılmıştır.

2. Modeller

2.1. Doğrusal regresyon modeli

i. bireyin bağımlı değişkeni Y_i, *p*-boyutlu açıklayıcı değişkenler vektörü $\boldsymbol{x_i} = (x_{i1}, \dots, x_{ip})$, *p*-boyutlu regresyon katsayıları vektörü $\boldsymbol{\beta}$ ve model hata terimi ε_i olmak üzere doğrusal regresyon eşitliği,

$$Y_i = \boldsymbol{x_i'\beta} + \varepsilon_i \tag{1}$$

biçiminde ifade edilir.

Eş. (1)' den de görüldüğü gibi, bağımlı değişken açıklayıcı değişkenlerin doğrusal fonksiyonudur. Bu modelde hata terimlerinin 0 ortalama ve sabit varyans ile birbirinden bağımsız olduğu ve normal dağıldıkları varsayılır, $\varepsilon_i \sim N(0, \sigma^2)$. Regresyon parametrelerinin tahmin edilmesinde genel olarak EKK yönteminden yararlanılmaktadır. Doğrusal regresyon, sağlık harcamalarının modellenmesinde sıklıkla kullanılsa da aykırı değerlere karşı duyarlı olması ve normallik varsayımının sağlanmaması durumunda etkin olmayan parametre tahminlerinin elde edilmesi açısından dezavantajları vardır [7].

i. bireyin toplam hasar tutarı S_i ile gösterilsin. Bu değişken doğrusal regresyon ile modellendiğinde aşağıdaki biçimde tanımlanır:

$$S_i = \eta_i + \varepsilon_i \tag{2}$$

Burada doğrusal kestirici, $\eta_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij}$ açıklayıcı değişkenler ve regresyon parametrelerinden oluşmaktadır. Eş. (2)'ye parametre tahminleri yerleştirildiğinde *i.* bireyin toplam hasar tutarının kestirimi aşağıdaki biçimde elde edilir [10]:

$$\hat{S}_i = \hat{\eta}_i = \hat{\beta}_0 + \sum_{j=1}^{p-1} \hat{\beta}_j x_{ij}$$

2.2. Lognormal model

Hasar tutarı verisi veya sağlık harcamaları verisi genellikle sağa çarpık dağılıma sahip olmaktadır. Çarpıklık sorununun giderilmesi ve daha simetrik dağılımın elde edilmesi için veriye çeşitli dönüşümler uygulanmaktadır. Bu dönüşümler arasında en sık kullanılanları logaritmik dönüşüm, kare-kök dönüşümü ve kuvvet dönüşümüdür. Logaritmik dönüşüm diğer dönüşümlere göre daha sık tercih edilmektedir. Logaritmik dönüşüm yapılarak elde edilen tahminlerin orijinal ölçeğe geri dönüştürülmesi için basit şekilde üstel alınması durumunda ortalama yerine medyanın tahmini elde edilmektedir [10].

Bağımlı değişkene logaritmik dönüşüm uygulanması durumunda model aşağıdaki biçimde gösterilir,

$$\log(Y_i) = \boldsymbol{x_i'\beta} + \varepsilon_i$$

Genellikle analizcinin amacı tutarların orijinal ölçekte tahmin edilmesidir. Bu nedenle aşağıdaki biçimde geri dönüşüm yapılmaktadır:

$$Y_i = exp(\boldsymbol{x_i'\beta} + \varepsilon_i)$$

Böylece koşullu beklenen değer,

$$E(Y_i|\boldsymbol{x_i}) = exp(\boldsymbol{x_i'\beta})E(exp(\varepsilon_i)|\boldsymbol{x_i})$$

şeklinde tanımlanır [12].

Sağlık harcamaları verisinde sağlık hizmetinden yararlanmayan bireylerin sıfır olan harcamalarının logaritmasının alınabilmesi için genellikle tüm tutarlara 1 eklenmektedir. Böylece *i.* bireyin toplam hasar tutarı için model,

$$log(S_i + 1) = \boldsymbol{x_i'\beta} + \varepsilon_i$$

biçiminde tanımlanır.

Eğer hata terimleri normal dağılıyor ise koşullu beklenen değer eşitliğinden yararlanılarak toplam hasar tutarının kestirimi,

$$\hat{S}_i = exp\left(\boldsymbol{x_i'}\widehat{\boldsymbol{\beta}} + 0.5\,\hat{\sigma}^2\right) - 1 \quad (3)$$

şeklinde elde edilir.

Eş. (3)'teki $\widehat{\boldsymbol{\beta}}$ ve $\hat{\sigma}^2$ değerleri, EKK yöntemi ile elde edilen tahmin değerleridir. Eğer hata terimleri normal dağılmıyor ise tahminler yanlı olur. Yanlılığın çözümü olarak Duan [6], model artık terimlerinin üstel bir fonksiyonu olarak hesaplanan smearing faktörünü önermiştir. Smearing faktörü kullanılarak toplam hasar tutarının kestirimi,

$$\hat{S}_i = exp\left(\boldsymbol{x_i'}\widehat{\boldsymbol{\beta}}\right)\hat{\phi} - 1 \quad (4)$$

biçimde elde edilir [10].

Eş. (4)'te $\hat{\phi}$ smearing faktörüdür ve aşağıdaki şekilde hesaplanmaktadır [6;10]:

$$\hat{\phi} = \frac{\Sigma_i\, exp(\hat{\epsilon}_i)}{K}, \quad \hat{\epsilon}_i = \log(S_i + 1) - \boldsymbol{x_i'}\widehat{\boldsymbol{\beta}} \quad (5)$$

Eş. (5)'te $\hat{\epsilon}_i$'ler model artık terimleri, K örneklem büyüklüğüdür. Smearing faktörü literatürde geri dönüşüm yöntemi olarak sıklıkla kullanılmaktadır [5;15]. Logaritmik dönüşüm yapılmış veride hata terimlerinde değişen varyans olması durumunda smearing faktörü ile yanlı kestirimler elde edilmektedir [12].

2.3. Genelleştirilmiş doğrusal model

Genelleştirilmiş doğrusal modeller, doğrusal regresyon modelinin genelleştirilmiş biçimidir. Genelleştirme iki temel özellikte özetlenebilir:

- Ortalamadan rastgele sapmaların dağılımı üstel aileden herhangi bir dağılım olabilir.
- Bağımlı değişkenin ortalaması, bağ fonksiyonuna göre belirlenen farklı ölçeklerde açıklayıcı değişkenler ile doğrusal olabilir [14].

GDM'de bağımlı değişkenin dağılımının, üstel ailenin bir üyesi olma varsayımı bulunmaktadır. Normal, Poisson, Binom, Gamma ve Ters Gauss dağılımları üstel aile üyesi olan dağılımlardır. Üstel aileye ait bir dağılıma sahip olan *Y* rastlantı değişkeninin olasılık (yoğunluk) fonksiyonu,

$$f(y;\theta,\phi) = exp\{\, a(\phi)^{-1}\,[y\theta - \psi(\theta)] + c(y,\phi)\} \tag{6}$$

biçiminde tanımlanır.

Eş. (6)'da θ kanonik parametre ve ϕ yayılım parametresi olarak adlandırılır. $\psi(.)$ ve *c(.)* fonksiyonları bilinen fonksiyonlardır. Genellikle $a(\phi) = \phi/w$ biçiminde tanımlanır. *w* bireye göre değişen ve önceden bilinen ağırlıklardır. ϕ biliniyor ise dağılım yalnızca θ kanonik parametreli üstel aile üyesidir. Bilinmiyor ise dağılım iki parametreli üstel aile üyesi olabilir veya olmayabilir.

Y rastlantı değişkeninin ortalama ve varyansı sırasıyla,

$$E(Y) = \mu = \psi'(\theta)$$

$$Var(Y) = \sigma^2 = a(\phi)\psi''(\theta)$$

biçiminde elde edilir [16].

Birey bazında tanımlama yapıldığında $i = 1,2,\dots,K$ olmak üzere Y_i, *i.* bireyin bağımlı değişkeni olarak tanımlanır ve Y_i'lerin birbirinden bağımsız olduğu varsayılır. Ortalama fonksiyonu, açıklayıcı değişkenlerin bir fonksiyonu olarak aşağıdaki biçimde tanımlanmaktadır:

$$\mu_i = h^{-1}(\boldsymbol{x_i'}\,\boldsymbol{\beta}\,) \tag{7}$$

Eş.(7)'de $\boldsymbol{x_i} = \left(x_{i1}, x_{i2}, \dots, x_{ip}\,\right)'$ *i.* bireyin *p*-boyutlu açıklayıcı değişkenler vektörü, $\boldsymbol{\beta}$ regresyon parametrelerinin katsayılar vektörüdür. *h(.)* fonksiyonu bağ fonksiyonu olarak tanımlanmakta olup, bağımlı değişkenin ortalaması ile doğrusal kestirici arasındaki ilişkiyi belirler. Sağlık harcamaları verisi ve sigorta verisinin modellenmesinde genellikle logaritmik bağ fonksiyonu tercih edilmektedir. Bu bağ fonksiyonu ile açıklayıcı değişkenlerin ortalama üzerinde çarpımsal etkisi oluşur. GDM'lerde tahmin sonuçlarının orijinal ölçekte ifade edilmesi için geri dönüşüme gerek olmaması, GDM'lerin lognormal modele göre önemli avantajıdır.

Doğrusal kestirici, $\eta_i = \boldsymbol{x_i'}\,\boldsymbol{\beta}$ olmak üzere GDM'de logaritmik bağ fonksiyonu kullanılması durumunda *i.* bireyin beklenen toplam hasar tutarının kestiricisi,

$$\hat{S}_i = exp(\hat{\eta}_i)$$

biçiminde elde edilir.

2.4. İki-kısım model

Lognormal model ve GDM gibi tek-kısım modeller, bağımlı değişkendeki yüksek sayıda sıfır değerlerini göz ardı edip, sağa çarpık dağılımlı pozitif harcamaların modellenmesine odaklanır. Oysa hasar tutarı verisi veya sağlık harcamaları verisinde gözlenen yüksek sayıdaki sıfır değerler hizmet kullanım eğilimi açısından önemli bilgiler içermektedir. İki-kısım modeller, veride gözlenen sıfırlar ile bireylerin sağlık hizmetinden yararlanması durumunda oluşan harcamalarının ayrı kısımlarda modellenmesine olanak vermektedir.

Aitchison [2], sıfır noktasında nokta yoğunluğuna sahip olan pozitif rastlantı değişkeninin modellenmesi fikrini ilk olarak ortaya çıkarmış ve bu karma dağılımın ortalama ve varyansını tahmin etmiştir.

S_i, *i.* bireyin toplam hasar tutarı olmak üzere,

$$I_i = \begin{cases} 1, & S_i > 0 \\ 0, & S_i = 0 \end{cases}$$

biçiminde indikatör tanımlansın. *i.* bireyin toplam hasar tutarının tahmin edilmesi için bireyin hasar bildirme olasılığı ile beklenen hasar tutarı aşağıdaki şekilde çarpılır:

$$E(S_i|\boldsymbol{x_i}) = P(I_i = 1|\boldsymbol{x_i})E(S_i|\boldsymbol{x_i}, I_i = 1)$$

$p_i = P(I_i = 1|\boldsymbol{x_i})$ pozitif hasar olasılığıdır. p_i olasılığı lojistik regresyon ile aşağıdaki biçimde tahmin edilebilir,

$$p_i = \frac{1}{1+e^{-\boldsymbol{x_i'}\boldsymbol{\beta}}} = \frac{1}{1+e^{-\varsigma_i}} = \frac{e^{\varsigma_i}}{1+e^{\varsigma_i}} \tag{8}$$

Eş. (8)'de ς_i lojistik regresyonun sistematik bileşenidir [13].

Bireylerin pozitif toplam hasar tutarlarının lognormal model ile modellendiği, geri dönüşümde smearing faktörünün kullanıldığı ve bireylerin hasar yapma olasılığının lojistik regresyon ile modellendiği durumda *i.* bireyin toplam hasar tutarının kestiricisi,

$$\begin{aligned} \hat{S}_i &= E\left(S_i|\hat{\eta}_i, \hat{\phi}, \hat{\varsigma}_i\right) \\ &= E\left(S_i|S_i > 0, \hat{\eta}_i, \hat{\phi}\right)P(S_i > 0|\hat{\varsigma}_i) \\ &= exp(\hat{\eta}_i)\hat{\phi}\frac{1}{1+e^{-\hat{\varsigma}_i}} \end{aligned}$$

biçiminde elde edilir.

Benzer şekilde pozitif hasar tutarlarının logaritmik bağ fonksiyonlu GDM ile modellenmesi durumunda *i.* bireyin toplam hasar tutarının kestiricisi aşağıdaki biçimde elde edilir [10]:

$$\begin{aligned} \hat{S}_i &= E(S_i|\hat{\eta}_i, \hat{\varsigma}_i) \\ &= E(S_i|S_i > 0, \hat{\eta}_i)P(S_i > 0|\hat{\varsigma}_i) \\ &= exp(\hat{\eta}_i)\frac{1}{1+e^{-\hat{\varsigma}_i}} \end{aligned}$$

3. Uygulama

3.1. Veri hakkında

Uygulama çalışmasında kullanılan veri seti, Türkiye'de hizmet gösteren bir sigorta şirketinden alınan özel sağlık sigortası verisidir. Tüm poliçeler bireysel poliçe olup grup poliçe bilgisi kullanılmamıştır. Hasar tutarları, bildirilen hasar karşılığında sigorta şirketinin poliçe sahibine ödediği tutarlardan oluşmakta, meydana gelmiş ancak bildirilmemiş hasar bilgisini ve reasürans uygulamasını içermemektedir. Ödenen hasar tutarlarında muafiyet uygulaması göz ardı edilmiştir. Çalışmada kullanılan veri, 18-70 yaşları arasındaki bireylerin bilgisini içermektedir. Tüm poliçeler bir yıl boyunca yürürlükte kalmıştır. Poliçe sahibinin kendisine ait bilgiler kullanılmış, bağımlılarına ilişkin bilgi kullanılmamıştır.

Modelleme kısmında kullanılan veri seti 2010 yılında başlamış veya son yenilemesi yapılmış 21.496 adet poliçe bilgisinden; model doğrulamada kullanılan veri seti ise 2011 yılında başlamış veya son yenilemesi yapılmış 22.057 adet poliçe bilgisinden oluşmaktadır. Uygulama çalışmasında, 2010 yılı verisi ile modelleme yapılmış, bu kısımdan elde edilen parametre tahminleri kullanılarak 2011 yılı poliçe sahiplerinin toplam hasar tutarının kestirimi yapılmıştır.

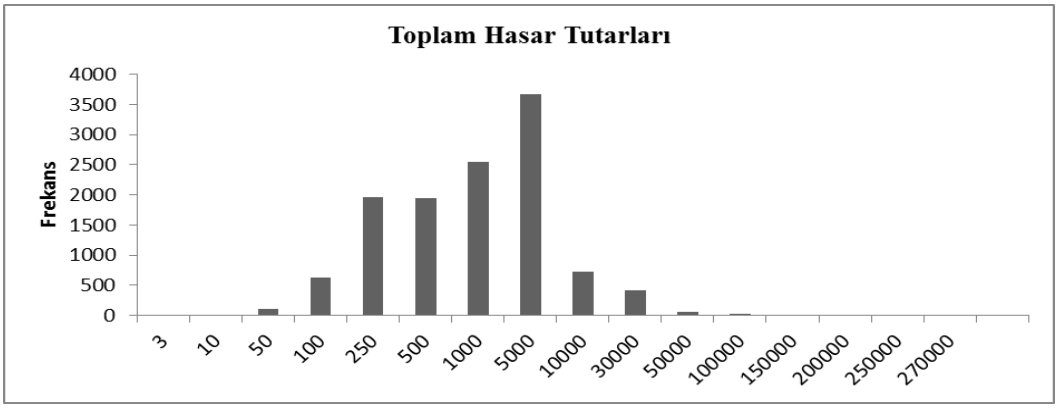
Çalışmanın amacı, poliçe yılı başında çeşitli özellikleri bilinen poliçe sahibinin yıl içerisinde getireceği toplam hasar tutarının kestirimi olması nedeniyle poliçe başlangıç tarihinde bilinen açıklayıcı değişkenler kullanılmıştır. Kullanılan açıklayıcı değişkenler; bireyin yaşı, cinsiyeti, medeni durumu, yaşadığı şehir ve satın aldığı sağlık sigortası ürününün paket grubudur. Yaş dışında diğer değişkenler kategoriktir. Şehirler poliçe sayılarına göre gruplandırılmıştır. İstanbul, Ankara ve İzmir illerinde satın alınan poliçe sayısı diğer illere göre belirgin şekilde yüksek olduğundan bu iller ayrı kategori olarak sınıflandırılmıştır. Diğer şehirler de poliçe sayılarına göre iki kategoride gruplandırılmıştır. Ürün paketleri çeşitli teminat tipleri ve limitlere göre farklılık göstermektedir. Ürün paketleri, ekonomik paket grubunda olup olmaması ve yatarak tedavinin yanında ayakta tedavi teminatı sağlayıp sağlamaması özelliklerine göre temel olarak dört kategoride gruplandırılmıştır. Nitel açıklayıcı değişkenler kategorileştirilirken temel seviyedeki (0) gözlem sayısının seyrek olmaması esas alınmıştır. Kategorik açıklayıcı değişkenlerin tanımları Çizelge 1'de verilmiştir.

Çizelge 1. Kategorik açıklayıcı değişkenler

Faktör	Kategori	Tanım
Şehir	Diğer	4
	Kocaeli, Bursa, Muğla, Antalya, Adana, Tekirdağ	3
	İzmir	2
	Ankara	1
	İstanbul	0
Paket	Eko, sadece yatarak	3
	Eko, yatarak+ayakta	2
	Eko olmayan, sadece yatarak	1
	Eko olmayan, yatarak+ayakta	0
Cinsiyet	Kadın	1
	Erkek	0
Medeni durum	Evli	1
	Bekar veya dul	0

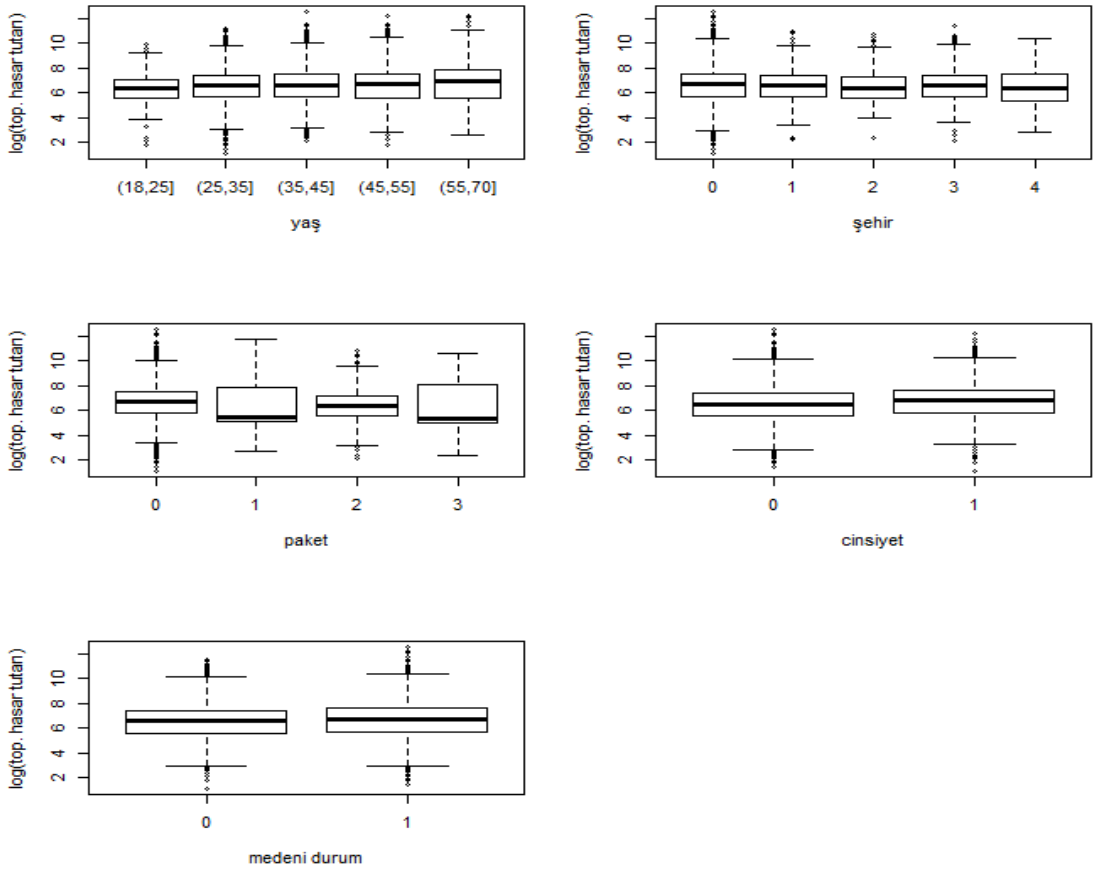
3.2. Modelleme

2010 yılı poliçe sahiplerinin bir yıllık poliçe döneminde sigorta şirketine bildirdikleri hasarları karşılığında şirket tarafından ödenen toplam hasar tutarlarının (TL) histogram grafiği Şekil 1'de verilmiştir.



Şekil 1. 2010 poliçe yılı-toplam hasar tutarlarının histogram grafiği

Bireylerin pozitif toplam hasar tutarları 3 TL – 267.272 TL aralığında değer almaktadır. Tutarların ortalaması 2.231 ve medyanı 759 olmak üzere Şekil 1'den de görüldüğü gibi toplam hasar tutarlarının dağılımı sağa çarpıktır. Sağa çarpık dağılım nedeniyle iki-kısım modellerde bireylerin pozitif toplam hasar tutarları gamma ve lognormal dağılım varsayımı altında modellenmiştir. Gamma dağılımı üstel ailenin bir üyesi olduğundan modellemede GDM kullanılmıştır. Lognormal modelde toplam hasar tutarlarına logaritmik dönüşüm uygulanmış, dönüşüm yapılan tutarlar doğrusal regresyon modeli ile modellenmiştir. Logaritmik dönüşüm yapılmış toplam hasar tutarlarının açıklayıcı değişkenlere göre kutu grafikleri Şekil 2'de verilmiştir:



Şekil 2. Logaritmik dönüşüm yapılan toplam hasar tutarlarının açıklayıcı değişkenlere göre kutu grafikleri

Şekil 2'ye bakıldığında yaş arttıkça bireylerin toplam hasar tutarının arttığı görülmektedir. Şehirlere bakıldığında İstanbul'da yaşayan poliçe sahiplerinin toplam hasar tutarlarının diğer şehirlerde yaşayan bireylerin toplam hasar tutarından yüksek olduğu görülmektedir. Ürün paketlerinin kategorilerine bakıldığında, yatarak tedavi teminatına ek olarak ayakta tedavi teminatı sağlayan ürün paketlerini satın alan poliçe sahiplerinin toplam hasar tutarlarına ilişkin medyan değerlerinin, yalnızca yatarak tedavi teminatı veren ürün paketlerini satın alan poliçe sahiplerinin toplam hasar tutarlarına ilişkin medyan değerlerinden yüksek olduğu görülmektedir. Cinsiyette kadın bireylerin toplam hasar tutarları erkek bireylere göre; medeni durumda ise evli bireylerin toplam hasar tutarları bekar veya dul bireylere göre küçük bir fark ile daha yüksektir.

Logaritmik dönüşüm uygulanmış pozitif toplam hasar tutarlarının doğrusal regresyon modeli ile modellenmesi, tutarların gamma dağılım varsayımı altında logaritmik bağ fonksiyonlu GDM ile modellenmesi ve bireylerin hasar yapma olasılıklarının lojistik regresyon ile modellenmesi sonucunda elde edilen tahmin sonuçları Çizelge 2'de verilmiştir.

Çizelge 2. İki-kısım model tahmin sonuçları

	Lognormal Model		Gamma GDM		Lojistik Regresyon	
Parametre	**Tahmin**	***p*-değeri**	**Tahmin**	***p*-değeri**	**Tahmin**	***p*-değeri**
Sabit terim	2,5939	$< 2x10^{-16}$*	6,7188	$< 2x10^{-16}$*	0,2780	$1,8x10^{-5}$*
Eko değil, sadece yatarak	-0,2132	$< 2x10^{-16}$*	0,2450	0,0033*	-1,5104	$< 2x10^{-16}$*
Eko, yatarak+ayakta	-0,1164	$1,3x10^{-10}$*	-0,3386	$3x10^{-5}$*	0,0353	0,4839
Eko, sadece yatarak	-0,2730	$9,8x10^{-5}$*	0,2710	0,3886	-1,8689	$< 2x10^{-16}$*
Ankara	-0,0405	0,0379*	-0,2757	0,0016*	-0,2949	$7,8x10^{-9}$*
İzmir	-0,0933	0,0002*	-0,3768	0,0009*	-0,2984	$3,5x10^{-6}$*
3. şehir kategori	-0,0421	0,0425*	-0,1195	0,1995	-0,1260	0,0224*
4. şehir kategori	-0,0747	0,0202*	-0,0393	0,7854	-0,1556	0,0617
Yaş	0,0058	$< 2x10^{-16}$*	0,0193	$7,1x10^{-16}$*	0,0095	$5,9x10^{-11}$*
Kadın	0,1413	$< 2x10^{-16}$*	0,1738	0,0004*	0,2100	$1,3x10^{-12}$*
Evli	0,0653	$2,4x10^{-9}$*	0,2008	$4,3x10^{-5}$*	-0,2502	$< 2x10^{-16}$*

*%95 güven düzeyinde istatistiksel olarak anlamlı

Lognormal modelde tahmin sonuçlarına bakıldığında tüm açıklayıcı değişkenlerin istatistiksel olarak anlamlı olduğu görülmektedir. Lognormal modele alternatif olarak kullanılan, verinin orijinal ölçekte modellendiği logaritmik bağ fonksiyonlu gamma GDM sonuçlarına göre, ekonomik sınıfta sadece yatarak tedavi teminatı sağlayan ürün paketi faktörü ile üçüncü ve dördüncü şehir faktörleri istatistiksel olarak anlamsızdır. Her iki model sonucuna göre, İstanbul'dan farklı şehirlerde yaşayan poliçe sahiplerinin toplam hasar tutarı İstanbul'da yaşayanlara göre daha düşüktür. Yaş arttıkça toplam hasar tutarı anlamlı bir biçimde artmaktadır. Kadın poliçe sahiplerinin toplam hasar tutarı erkek poliçe sahiplerine göre anlamlı biçimde daha yüksektir. Benzer şekilde evli bireylerin toplam hasar tutarı bekar veya dul bireylere göre anlamlı biçimde daha yüksektir.

Bireylerin hasar yapma olasılığının modellendiği lojistik regresyon sonuçlarına göre, sadece yatarak tedavi teminatı sağlayan paket faktörleri, dördüncü kategori dışında şehir faktörleri, kadın cinsiyet faktörü, evli olma medeni durum faktörü ve yaş istatistiksel olarak anlamlıdır.

Tek-kısım model olarak, sıfırların da dahil olduğu toplam hasar tutarı verisine doğrusal regresyon modeli ve lognormal model uygulanmıştır. Lognormal modelde logaritmik dönüşüm uygulanırken sıfır değerler nedeniyle her bir hasar tutarına 1 eklenmiştir.

3.3. Kestirim

Modellerin kestirim performansının karşılaştırılması için model doğrulama yöntemi kullanılmıştır. Aday modeller bir önceki kısımda verildiği gibi 2010 yılı verisine uygulanmış, bu kısımdan elde edilen parametre tahminleri kullanılarak 2011 yılı poliçe sahiplerinin toplam hasar tutarının nokta kestirimi yapılmış, bireylerin toplam hasar tutarının kestirim değeri ile gözlenen değeri arasındaki farktan yararlanılarak modellerin kestirim performansının ölçülmesi için aşağıdaki kriterler kullanılmıştır:

- Hata kareler ortalamasının karekökü: $\mathrm{RMSE} = \sqrt{\frac{1}{K}\sum_i \left(S_i - \hat{S}_i\right)^2}$
- Ortalama Mutlak Hata: $\mathrm{MAE} = \frac{1}{K}\sum_i \left|S_i - \hat{S}_i\right|$

Pozitif hasar tutarları için lognormal modelin kullanıldığı iki-kısım modelde kestirim değerleri elde edilirken smearing faktörü kullanılmıştır. Sıfırların da dahil olduğu toplam hasar tutarı verisine uygulanan tek-kısım lognormal modelde ise medyan kestiricisinin performansının ölçülmesi açısından smearing faktörü kullanılmamıştır.

Model doğrulama sonucunda elde edilen RMSE ve MAE değerleri Çizelge 3'te verilmiştir.

Çizelge 3. Model doğrulama sonuçları

Kestirimci Model	RMSE	MAE
İki-kısım model (Gamma GDM+Lojit model)	**5.724,492**	**1.708,739**
İki-kısım model (Log (y)-EKK+Lojit model)	5.757,946	1.718,519
Tek-kısım model (Doğrusal Regresyon)	**5.724,075**	1.711,366
Tek-kısım model (Lognormal Model (Log (y+1)-EKK))	5.893,389	**1.436,962**

Modellerin kestirim değerlerine ilişkin istatistikler Çizelge 4'te verilmiştir.

Çizelge 4. Toplam hasar tutarlarının kestirim değerlerine ilişkin istatistikler

Model	Min.	I. Çeyrek	Medyan	Ortalama	III. Çeyrek
İki-kısım model (Gamma GDM+Lojit model)	213	779	1.089	1.148	1.414
İki-kısım model (Log (y)-EKK +Lojit model)	116	455	1.070	1.069	1.557
Tek-kısım model (Doğrusal Regresyon)	-488	785	1.155	1.142	1.466
Tek-kısım model (Lognormal Model (Log (y+1)-EKK))	1,730	4,185	11,160	9,547	14,070

Çizelge 3'te verilen sonuçlara bakıldığında, tek-kısım modellerden doğrusal regresyonun RMSE kriterine göre; lognormal modelin MAE kriterine göre en düşük kestirim hatası verdiği görülmektedir. Ancak, Çizelge 4'te verilen kestirim değerlerinin istatistiklerine bakıldığında, normallik varsayımına dayanan doğrusal regresyondan negatif kestirim değerlerinin elde edildiği görülmektedir. Hasar tutarları için negatif kestirimlerin elde edilmesi bu modeli kullanışsız yapmaktadır. Lognormal modelden elde edilen kestirim değerlerine ilişkin istatistikler ise çok düşük değerlerdedir. Toplam hasar tutarlarının kestirim değerlerinin çok düşük olması, beklenen toplam hasar tutarlarına göre risk primi belirleyen sigorta şirketi açısından istenmeyen bir sonuçtur. Bu durum tek-kısım lognormal modeli de kullanışsız kılmaktadır. Sonuç olarak tek-kısım doğrusal regresyon modeli ve lognormal modelin toplam hasar tutarının kestirimi için uygun modeller olmadığı görülmüştür.

Tek-kısım modeller göz ardı edildiğinde, her iki kritere göre pozitif hasar tutarlarının gamma GDM ile modellendiği iki-kısım model, hasar tutarlarının lognormal model ile modellendiği iki-kısım modelden daha iyi performans göstermiştir.

4. Sonuçlar

Bu çalışmada, hayat-dışı sigorta branşında sigorta şirketinde hasar verisi olarak yalnızca bireylerin toplam hasar tutarına ilişkin bilgilerinin olduğu durum göz önüne alınarak poliçe dönemi başında çeşitli bilgileri bilinen poliçe sahiplerinin toplam hasar tutarının kestiriminin yapılması amaçlanmıştır. Kullanılan kestirim modelleri, sağlık harcamalarının modellenmesinde sıklıkla kullanılan tek-kısım modeller ve iki-kısım modellerdir.

Uygulama çalışmasında Türkiye'de faaliyet gösteren bir sigorta şirketinden alınan özel sağlık sigortası verisi kullanılmıştır. Modellerin kestirim performansının karşılaştırılmasında iki kriter kullanılmıştır; hata kareler ortalamasının karekökü ve ortalama mutlak hata. Her iki kritere göre tek-kısım modellerden elde edilen kestirim hataları iki-kısım modellere göre düşük olsa da kestirim değerlerinin istatistiklerine bakıldığında bu modellerin toplam hasar tutarının kestirimi veya prim belirlenmesi için uygun modeller olmadığı görülmüştür.

İki-kısım modellerde toplam hasar tutarları sağa çarpık dağılım nedeniyle lognormal ve gamma dağılımı varsayımı altında modellenmiştir. Lognormal modelin uygulanması için hasar tutarlarına logaritmik dönüşüm yapılmış, geri dönüşümde kestirim değerleri smearing faktörü ile çarpılmıştır. Ancak logaritmik ölçekte hata terimlerinde değişen varyans sorunu mevcut ise smearing tahmin edicisinin yanlılığa neden olduğu bilinmektedir. Bu nedenle veriye dönüşüm yapılması yerine verinin orijinal ölçekte GDM ile modellenmesi daha çok tercih edilmektedir. Bu çalışmada elde edilen sonuçlara göre, iki kısım modelde hasar tutarlarının gamma GDM ile modellendiği kestirim modeli her iki kritere göre tutarların lognormal model ile modellendiği kestirim modelinden daha iyi performans göstermiştir.

Kaynaklar

[1] A. Şentürk Acar, 2016, *Heterojenliğin sağlık sigortalarında toplam hasar modellerine etkisi*, Doktora tezi, Hacettepe Üniversitesi, Türkiye.

[2] J. Aitchison, 1955, On the distribution of a positive random variable having a discrete probability mass at the origin, *Journal of the American Statistical Association*, 50(271): 901–8.

[3] D. K. Blough, C. W. Madden, M. C. Hornbrook, 1999, Modeling risk using generalized linear models, *Journal of Health Economics*, 18(2):153–71.

[4] M. J. Brockman, T. S. Wright, 1992, Statistical motor rating: Making effective use of your data, *Journal of the Institute of Actuaries*, 119(03):457–543.

[5] M. B. Buntin, A. M. Zaslavsky, 2004, Too much ado about two-part models and transformation? Comparing methods of modeling medicare expenditures, *Journal of Health Economics* 23(3):525–42.

[6] N. Duan, 1983, A nonparametric smearing estimate : Method retransformation, *Journal of the American Statistical Association*, 78(383):605–10.

[7] N. Duan, W. G. Manning, C. N. Morris, J. P. Newhouse, 1983, A comparison of alternative models for the demand for medical care, *Journal of Business & Economic Statistics*, 1(2):115–26.

[8] E. W. Frees, J. Gao, M.A. Rosenberg, 2011, Predicting the frequency and amount of health care expenditures, *North American Actuarial Journal*, 15(3):377–92.

[9] E. W. Frees, R. A. Derrig, G. Meyers, 2014, *Predictive modeling applications in actuarial science*, 1st ed., Cambridge University Press.

[10] M. Griswold, G. Parmigiani, A. Potosky, J. Lipscomb, 2004, Analyzing health care costs: A comparison of statistical methods motivated by medicare colorectal cancer charges, *Biostatistics*, 1(1):1–23.

[11] S. Gschlößl, C. Czado, 2007, Spatial modelling of claim frequency and claim size in non-life insurance, *Scandinavian Actuarial Journa,l* 2007(3):202–25.

[12] A. M. Jones, 2010, *Models for health care*, HEDG Working Papers, 10/01, Department of Economics, University of York.

[13] P. Jong, Z. G. Heller, 2008, *Generalized linear models for insurance data*, London: Cambridge University Press.

[14] R. Kaas, M. Goovaerts, J. Dhaene, M. Denuit, 2008, *Modern actuarial risk theory using R*, Verlag Berlin Heidelberg: Springer.

[15] W. G. Manning, 1998, The logged dependent variable, heteroscedasticity and the retransformation problem, *Journal of Health Economics*, 17(3):283–95.

[16] P. McCullagh, J. A. Nelder, 1989, G*eneralized linear models*, Second Edi, London: Chapman and Hall.

[17] J. Mullahy, 1998, Much ado about two: Reconsidering retransformation and the two-part model in health econometrics, *Journal of Health Economics*, 17(3):247–81.

[18] P. Shi, X. Feng, A. Ivantsova, 2015, Dependent frequency – severity modeling of insurance claims, *Insurance: Mathematics and Economics*, 64:417–28.