Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics
Turkiye Klinikleri J Biostat. 2020;12(3):303-19

ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2020-75787

Cepten Sağlık Harcamasının Tahmininde Copas Testi ve İstatistiksel Daraltıcı (Shrinkage) Modellerin Uygulanması

Usage of Copas Test and Shrinkage Models to Predict Out-of-Pocket Health Expenditures

Songül ÇINAROĞLU[a]

[a]Hacettepe Üniversitesi İktisadi ve İdari Bilimler Fakültesi, Sağlık Yönetimi Bölümü, Ankara, TÜRKİYE

ÖZET Amaç: Sağlık harcaması değişkeninin, normal dağılım özelliği göstermeyerek aşırı derecede sağa çarpık olması doğrusallıktan ayrılmayı beraberinde getirerek, regresyon modelini zayıflatmaktadır. Copas testi ile istatistiksel daraltıcı (Shrinkage) modellerin kullanılması yolu ile tahmin modeli performansının iyileştirilmesi mümkündür. Bu çalışmada, cepten sağlık harcamasını tahmin etmek amacıyla Copas testi ve istatistiksel daraltıcı modelleri uygulanarak model performansları karşılaştırılmıştır. **Gereç ve Yöntemler:** Türkiye İstatistik Kurumu 2012 yılı Hane halkı Bütçe Anketi verileri kullanılmıştır. Cepten sağlık harcamasının tahmininde ödeme gücü değişkeni ile hane reisinin cinsiyet, 65 yaş üzerinde olma, sağlık sigortası ve kır ya da kentte yerleşim durumu değişkenleri kullanılmıştır. Eğitim veri setinin orijinal veri setini %50-95 oranlarında temsil ettiği 10 farklı eğitim ve test veri setleri oluşturulmuş ve Copas testi kullanılarak, en iyi performans sergileyen veri seti seçilmiştir. En küçük kareler (EKK) regresyonu ile istatistiksel daraltıcı olarak da bilinen Lasso ve Ridge regresyon teknikleri uygulanmıştır. **Bulgular:** EKK modelinin en düşük ortalama karesel hataya sahip olduğu ve en iyi performansı sergilediği görülmüştür. Cepten sağlık harcamasını tahmin etmede en etkili değişken ödeme gücü olup, Lasso regresyonun, Ridge regresyona göre daha iyi tahmin gücüne sahip olduğu görülmüştür. Lasso regresyon için optimal büzülme parametre değerinin eğitim veri seti için λ=0,0158; test veri seti için ise λ=0,0630 olduğu bulunmuştur. **Sonuç:** Copas testinin sağlık harcaması gibi modellenmesi zor değişkenler için en iyi model performansının araştırılmasında yararlı bir teknik olduğu, EKK ve Lasso regresyonun iyi tahmin performansına sahip olduğu ve ödeme gücü değişkeninin modele en fazla katkı sağlayan değişken olduğu belirlenmiştir. Farklı regresyon modelleri kullanılarak yapılacak incelemeler ile çalışma bulgularının geliştirilmesi tavsiye edilmektedir.

Anahtar kelimeler: Cepten sağlık harcaması; Copas testi; EKK regresyonu; Lasso regresyon; Ridge regresyon

ABSTRACT Objective: Not normal and extremely positively skewed distribution of health expenditures causes departure from normality and poor regression performance. It is possible to improve prediction performance with the usage of Copas test and Shrinkage models. In this study Copas test and Shrinkage models are used and compared to predict out-of-pocket health expenditures. **Material and Methods:** TurkStat Household Budget Survey for the year 2012 was used. Capacity to pay, which is a kind of adjusted income variable, gender of head of household, being older than 65 years old, health insurance and rural or urban settlement variables are used to predict households out-of-pocket health expenditures. Ten different training and test data sets were created, in which the training data set represented 50% to 95% of the original data set, and the data set with the best performance was selected by applying the Copas test. Least squares regression (LSR) and Lasso and Ridge regression techniques are applied which are also known as Shrinkage methods. **Results:** It is seen that LSC model had the lowest mean square error and had superior performance. The most effective variable in predicting out-of-pocket health expenditure is capacity to pay, and Lasso regression was found to have better predictive power than Ridge regression. It is seen that optimal Shrinkage parameter for Lasso regression for training data is λ=0.0158 and λ=0.0630 for test data. **Conclusion:** It has been determined that the Copas test is a useful technique in investigating the best model performance for variables that are difficult to model, such as health expenditure, LSC regression performs better than the Shrinkage models, and capacity to pay is the variable that contributes most into the model. It is recommended to develop study findings by examining different regression models.

Keywords: Out-of-pocket health expenditure; Copas test; LSR regression; Lasso regression; Ridge regression

Correspondence: Songül ÇINAROĞLU
Hacettepe Üniversitesi İktisadi ve İdari Bilimler Fakültesi, Sağlık Yönetimi Bölümü, Ankara, TÜRKİYE/TURKEY
E-mail: cinaroglus@hacettepe.edu.tr

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 18 May 2020 ***Received in revised form:*** 16 Sep 2020 ***Accepted:*** 17 Sep 2020 ***Available online:*** 21 Dec 2020

2146-8877 / Copyright © 2020 by Türkiye Klinikleri. This in an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Herhangi bir ülkede sağlık harcamalarının seviyesi, ülke genelinde gelişmişlik ve refah düzeyinin önemli bir göstergesidir. Bu nedenle sağlık harcamalarının belirleyicilerinin neler olduğu konusu dikkatle ele alınan bir konudur. Bu konu ile ilgili yapılan çalışmalarda sağlık harcamalarının belirleyicisi olarak farklı göstergeler dikkate alınmakla birlikte ülkelerin gelişmişlik seviyesi belirleyici rol oynamaktadır.[1,2] Sağlık harcamaları içerisinde büyük bir yer tutan cepten sağlık harcamaları, hane halkları tarafından doğrudan sağlık amaçlı yapılan harcamaları içermektedir.[3] Sağlık harcamasına ait dağılımın normal dağılım özelliği göstermeyerek aşırı sağa çarpık olması, sağlık harcamalarının modellendiği çalışmalarda model performansını bozucu bir etki yaratmaktadır. Bu nedenle doğrusallığı yakalamak için farklı dönüşüm yaklaşımlarının uygulanması, veri setinin farklı boyutlara ayrılması, farklı regresyon yöntemlerinin karşılaştırılması gibi alternatif yaklaşımlar uygulanabilmektedir. Bu çalışmada ise cepten sağlık harcaması değişkeninin tahmininde, bu alternatif dönüşüm ve regresyon modellerinin uygulanması ve karşılaştırmalı olarak incelenmesi hedeflenmiştir.

Cepten yapılan sağlık harcamalarından yola çıkılarak bulunan bir harcama değeri ise ödeme gücüdür. Ödeme gücü kendi eş değer hane büyüklüğüne göre belirlenen yoksulluk sınırı, gıda harcamasından küçük ve eşit olan hanelerde yoksulluk sınırının dışında kalan harcama değeri, kendi eş değer hane büyüklüğüne göre belirlenen yoksulluk sınırı, gıda harcamasından büyük olan hanelerde ise gıdanın dışında kalan harcama değeri olarak ifade edilmektedir.[4] Xu tarafından geliştirilen metodolojiye uygun olarak ödeme gücünün hesaplanmasında öncelikle eş değer hane halkı büyüklüğü (*eqsize*) belirlenmektedir.[5] Eş değer hane halkı büyüklüğü, hane halklarının erişkin-çocuk bileşimindeki farklılıkları dikkate alarak, farklı büyüklük ve birleşimdeki hane halkları arasında daha doğru kıyaslamalar yapılmasına olanak sağlayan katsayılar, eş değerlik ölçekleridir. Bu çalışmada eş değerlik ölçeği Eşitlik 1'de görülen aşağıdaki formülasyona göre Xu tarafından geliştirilen yaklaşım esas alınarak, beta katsayısı (β) 0,56 olarak kullanılmıştır.[5] Bu katsayı, aralarında Türkiye'nin de yer aldığı 59 ülke üzerinde yapılan modelleme çalışması ile elde edilmiştir. Bu katsayı, hane halkını oluşturan kişi sayısı arttıkça gıda harcamasının artması durumu ile başa çıkabilmek için geliştirilmiş bir tür ayarlama yaklaşımıdır. Bu sayede gıda harcaması ile hane halkı kişi sayılarına ait oransal artış miktarları dengelenmiş olur.[5]

$$eqsize_h = hhsize_h^{\beta} \quad \textbf{(1)}$$

Bu eşitlikte "$hhsize_h$" gerçek hane halkı büyüklüğünü, "β" ise Xu tarafından 59 ülkede hane halkı çalışması verisine dayanılarak hesaplanan 0,56 değerini göstermektedir.[5] Eş değer hane halkı büyüklüğü hesaplandıktan sonra yoksulluk sınırının (*poverty line=pl*) bulunması gerekmektedir.[5]

Eşitlik 2'de görüldüğü üzere, gıda harcamasının toplam tüketim harcaması içerisindeki payına göre sıralı %45-55 arasında kalan hanelerin, eş değer fert başına gıda harcamalarının ortalaması yoksulluk sınırı (*pl*) olarak elde edilmektedir. Burada;

$$pl = \frac{\sum w_h * eqfood_h}{\sum w_h} \quad \textbf{(2)}$$

w_h Faktör (kitle genişletme katsayısı-nüfus ağırlığı)

eqfood Eş değer fert başına gıda harcaması (*food/eqsize*)

foodexp Gıda harcamasının toplam harcamaya oranı (*food/exp*)

Daha sonra eş değer hane büyüklüklerine göre geçinme sınırı (subsistence expenditure, *se*) tespit edilmektedir.

$$se_h = pl * eqsize_h \quad \textbf{(3)}$$

Eğer $exp_h < se_h$ ise $poor_h$=1

Eğer $exp_h \geq se_h$ ise $poor_h$=0

Buna göre hanenin harcaması, kendi eş değer hane büyüklüğüne göre belirlenen geçinme sınırından küçük olan haneler "yoksul" olarak isimlendirilmektedir.[5]

Hanenin ödeme gücü ya da kapasitesi (household's capacity to pay-ctp) geçinme sınırı dışında kalan harcaması olarak kabul edilmektedir. Ödeme gücü ya da kapasitesi;[5]

- Kendi eş değer hane büyüklüğüne göre belirlenen yoksulluk sınırı, gıda harcamasından küçük ve eşit olan hanelerde yoksulluk sınırı dışında kalan harcama değeri,
- Kendi eş değer hane büyüklüğüne göre belirlenen yoksulluk sınırı, gıda harcamasından büyük olan hanelerde ise gıdanın dışında kalan harcama değeri olarak tanımlanmaktadır.

Sağlık harcamalarının konu edinildiği modellerde karşılaşılan önemli bir güçlük istatistiksel modelin verinin temel yapısını yakalayamaması durumunda ortaya çıkan bağımlı ve bağımsız değişkenler arasındaki ilişkide doğrusallıktan ayrılma ve model performansında düşüştür.[6-9] Bunun yansı sıra sağlık harcamalarının ayrıntılı olarak incelendiği çalışmalarda karşılaşılan temel sorunlardan birisi sağlık harcaması dağılımının aşırı sağa çarpık olmasıdır.[7]

Sağlık harcamaları ile ilgili analizlerde sorun yaratan bu normal dağılımdan ayrılma durumu ile başa çıkabilmek için farklı dönüşüm yaklaşımları uygulanmaktadır. Bu sayede sağlık harcaması dağılımının, normal dağılıma yaklaştırılması mümkün olabilmektedir. Bu yaklaşımlar içerisinde en fazla logaritmik dönüşümün tercih edildiği görülmektedir.[7] Bilger ve Manning'e göre farklı dönüşüm yaklaşımları arasında sağlık ekonomistleri arasında yaygınlıkla kabul edilen tek bir görüş bulunmamakla birlikte, dağılımın normalleştirilmesi amacıyla deneme yanılma yönteminin uygulanması ve farklı yöntemler sabırla uygulanarak verideki değişikliklerin izlenmesi tavsiye edilmektedir.[10,11] Normallik dönüşümleri sağlandıktan sonra neden-sonuç ilişkisi bulmaya yönelik oluşturulacak regresyon modellerinde, harcama gibi modellenmesi zor değişkenlerin kullanılması durumunda, standart regresyon analizi ile elde edilen tahminlerden istenilen performansın elde edilememesi durumu ile karşılaşılmaktadır.[12] Bu sorunun çözümü için cezalı regresyon (penalized regression) yöntemlerinden faydalanılmaktadır. Bu teknikler arasında sayılan Least Absolute Shrinkage and Selection (Lasso) ve Ridge regresyon, farklı cezalandırma kullanırlar. Tibshirani tarafından belirtildiği üzere Lasso:[13]

$\sum|\beta_j| \leq$ t koşulu ile $\min_\beta (y - X\beta)^T (y - X\beta)$ olarak verilmiştir. Burada $t \geq 0$ ayarlama parametresi (tunning parameter) olarak kullanılmaktadır.

Lasso, EKK tahmin edicisi $\hat{\beta}_{EKK}$'yi sıfıra büzebilir ve böylece bazı *j* değerleri için $\hat{\beta}_j = 0$ olabilir.

Ridge regresyonun amaç fonksiyonu ise:

$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{m} x_{ij}\,\beta_j)^2 + \lambda_R \sum_{j=1}^{m} \beta_j^2 \rightarrow min$ 'dir. **(4)**

Problemin çözümü

$\hat{\beta}_{RDG} = (X^T X + \lambda_R I)^{-1} X^T y$'dir. **(5)**

Varyansı $\lambda_R>0$ için $\hat{\beta}_{EKK}$'den daha küçüktür.

$$Var\ (\hat{\beta}_{RDG}) = (X^TX + \lambda_R I)^{-1}X^TX(X^TX + \lambda_R I)^{-1}\sigma^2 \quad \textbf{(6)}$$

$$< Var\left(\hat{\beta}_{EKK}\right) = \sigma^2\ (X^TX)^{-1}$$

λ_L ve λ_R parametreleri büzülme miktarlarını kontrol eder ve 0 ya da daha büyük değerler seçilmelidir. Eğer parametre 0 ise EKK gibi Lasso ve Ridge regresyonun sonuçları da aynıdır.[13,14]

Ridge regresyonda büzülme etkisini, 2 açıklayıcı değişkenli bir doğrusal regresyon problemini düşünerek örneklendirecek olursak, Ridge tahmin edicinin varyansı:

$$Var\ (\hat{\beta}_{RDG}) = \begin{pmatrix} 1+\lambda & r \\ r & 1+\lambda \end{pmatrix}^{-1} \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \begin{pmatrix} 1+\lambda & r \\ r & 1+\lambda \end{pmatrix}^{-1} \quad \textbf{(7)}$$

yanlılık

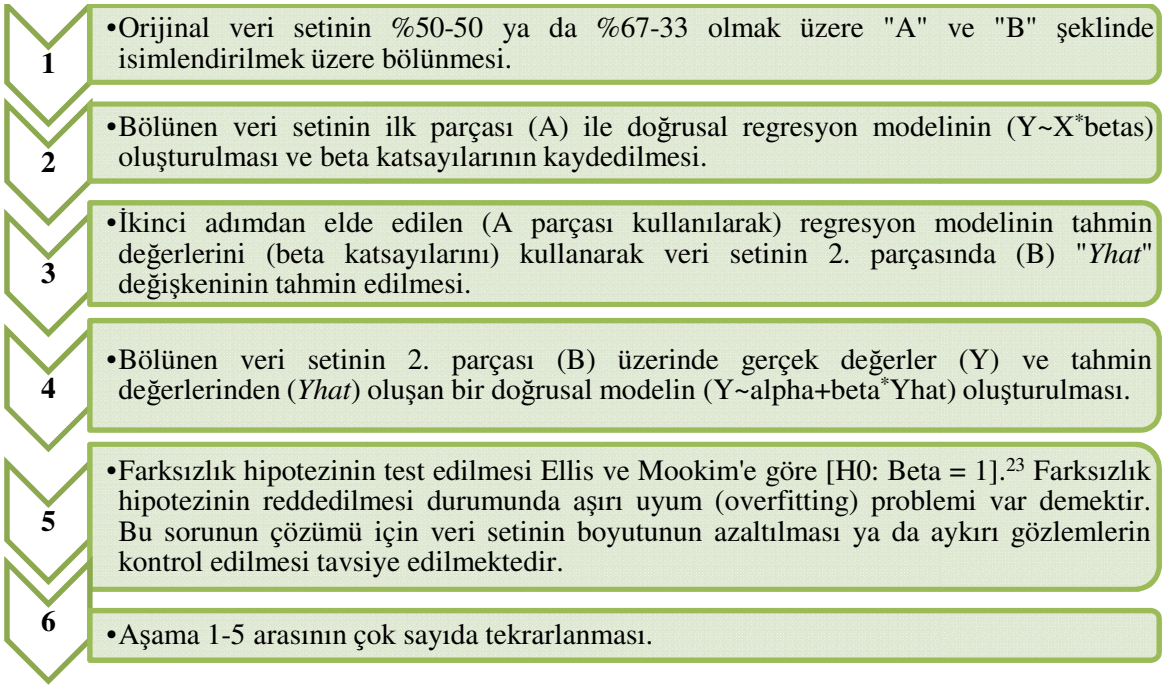
$$Bias\ (\hat{\beta}_{RDG}) = \begin{pmatrix} 1+\lambda & r \\ r & 1+\lambda \end{pmatrix}^{-1} \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} - \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \quad \textbf{(8)}$$

ve Hata Kareler Ortalaması - Mean Squared Error (MSE)

$MSE = Var\ (\hat{\beta}_j) + [Bias\ (\hat{\beta}_j)]^2$ olarak ifade edilir. **(9)**

Sağlık harcamalarının tahmin edilmesine yönelik oluşturulan modellerde normallik dönüşümleri ve farklı regresyon teknikleri kullanılarak, bağımlı ve bağımsız değişkenler arasındaki ilişkide doğrusallığa yaklaşmak ve bağımlı değişken değerlerine yönelik daha güçlü kestirimler yapmak mümkündür.[7,10,15] Doğrusal olmayan modellerde ölçüm, genellikle bağımlı değişkenin ait olduğu ölçüm sınıfından farklı bir ölçekte değerlendirilmesini gerektirmektedir. Bu noktada, logaritmik ya da Box-Cox (BC) dönüşümü gibi farklı dönüşümlerin uygulanması tavsiye edilmektedir.[16] Sağlık ekonomisi alanında uzman araştırmacılar, sağlıkta politika belirleyicilerin her ne kadar sağlık harcamalarının modellendiği çalışmalarda logaritmik dönüşüm uygulanmış sağlık harcaması değişkeni ile ilgilenmediklerini belirtseler de daha iyi tahmin sonuçlarına erişebilmek için dönüşüm uygulamak iyi bir strateji olarak önerilmektedir.[10]

Sağlık harcamalarının modellendiği çalışmalarda cezalı regresyon yöntemleri tahminleri (örneğin Ridge regresyon) dengelediği için bir tür "büzülme modeli (Shrinkage)" olarak isimlendirilmektedir.[17] Ayrıca bir örneklemden elde edilen parametre tahminleri aynı popülasyondan yeni bir örneklemde tahminler yapmak için kullanıldığında, aşırı uyum (overfitting) sorununun oluşmaması durumunda gerçek sonuçlar ile tahminlerin grafiği 45 derecelik doğru üzerinde yer almalıdır. Kırk beş derecelik doğrudan ayrılmalar ise "Shrinkage" olarak isimlendirilmektedir.[10] Başka bir deyişle, bağımlı ve bağımsız değişken arasındaki ilişkide aşırı uyumun bulunmaması durumunda, Shrinkage alanı olarak ifade edilen regresyon doğrusu eğiminin, 45 derecelik bir açı özelliği gösterdiği görülmektedir.[10] Shrinkage alanı Larson tarafından öncülük edildiği üzere çapraz geçerlilik teknikleri kullanılarak ölçülebilmektedir.[18] Zaman içerisinde Copas tarafından Shrinkage alanının ölçülmesi için önerilen yaklaşım popülerlik kazanmış ve sağlık bilimleri (örneğin Harrell ve ark. 1996; Harrell 2001) ve sağlık ekonometrisi (örneğin Blough ve ark. 1999; Bilger ve Manning 2015) alanlarında uzman araştırmacılar tarafından "*Copas aşırı uyum testi*" olarak isimlendirilerek uygulanmıştır.[10,19-22] Aşağıda, Copas Shrinkage ve aşırı uyum testinin aşamaları görülmektedir. Bu çalışmada ise sağlık harcamasının modellendiği birçoklu regresyon modeli oluşturularak, Shrinkage alanının ölçümünde Şekil 1'de aşamaları gösterilen Copas testinden faydalanılmıştır.



ŞEKİL 1: Copas testinin aşamaları.[19,23,24]

Bu çalışmanın son aşamasında ise istatistiksel daraltıcı (Shrinkage) modeller için kritik bir düzeltme parametresi olarak kabul edilen lambda (λ) değerini optimize etmek için Lasso regresyonda çapraz geçerlilik uygulanmıştır.[25,26] Bunun için R programında "*glmnet*" paketi kullanılmıştır.[27]

Çapraz geçerlilik, Lasso'da büzülme (Shrinkage) derecesine karar vermede önemli bir kriterdir. Çapraz geçerlilikte hedef tahmin gücü en yüksek modeli bulabilmektir.[28] Bu süreçte öncelikle veri seti rastgele V sayıda ve her parçada yaklaşık m kişi bulunacak biçimde kabaca eşit parçalara ayrılmaktadır. Daha sonra, her bir parça x_v ve y_v $(v \in \{1, \dots, V\})$, ve kalan $V-1$ parçalar da x_{-v} ve y_{-v} olarak isimlendirilmiş olan doğrulama (validation) verisi olarak kaydedilmektedir. Model, belirlenen bir λ değeri için eğitim veri setine uygulanmakta, daha sonra doğrulama veri setinde y_v için $\hat{y}_v(\lambda)$ tahminini elde etmek için kullanılmaktadır. Modelin ortalama tahmin yeteneği şu şekilde değerlendirilebilmektedir:[29]

$$P_{CV}(\lambda) = \frac{1}{V} P(y_v, \hat{y}_v(\lambda)) \quad \textbf{(10)}$$

Eşitlik 10'da $P(y_v, \hat{y}_v(\lambda))$ fonksiyonu tahmin doğruluğunun net bir ölçüsü olarak kabul edilmektedir ve MSE tahmin doğruluğunun bir ölçüsü olarak şu şekilde tanımlanmaktadır:[29]

$$P\left(y_v, \hat{y}_v(\lambda)\right) = \frac{1}{m} \left(y_v - \hat{y}_v(\lambda)\right)^T \left(y_v - \hat{y}_v(\lambda)\right) \quad \textbf{(11)}$$

Eşitlik 11'de $\hat{\lambda}$ olarak tanımlanan ve Eşitlik 10'da ortalama tahmin hatasını minimize eden optimal büzülme (Shrinkage) değerinin elde edilmesi hedeflenmektedir.[29]

Optimal λ parametresi belirlenirken gözden kaçırılmaması gereken nokta, çok küçük parametre değerlerinin aşırı uyum sorununa yol açabileceği, çok büyük λ parametre değerinin ise uyumsuzluk sorunu ortaya koyabileceğidir. Her 2 durumda da yüksek tahmin hatası elde etmek mümkündür.[29]

Çalışmanın orijinal katkıları arasında (*i*), sağlık harcamasının tahmin edilmesinde Xu tarafından geliştirilen metodolojiye uygun olarak hesaplanmış ve hane halkı seviyesinde bir tür ayarlanmış değişken olan ödeme gücü değişkeni ile cinsiyet, 65 yaş üzerinde olma, sağlık sigortasının bulunması, kır ya da kentte ikamet etmek bağımsız değişkenleri kullanılarak birçoklu regresyon modeli oluşturulacaktır.[5] (*ii*) Orijinal

veri seti farklı büyüklüklere ayrılacak ve eğitim ve test veri setlerinin performansları ayrı ayrı incelenecektir. (*iii*) Copas tarafından önerilen ve zaman içerisinde sağlık ekonometrisi alanında popülerlik kazanan yaklaşım ilk kez bu çalışmada, Türkiye'de sağlık harcamalarının konu edinildiği ve hane halkı düzeyindeki bir veri seti üzerinde kullanılacaktır.[19] (*iv*) Cepten sağlık harcamasını tahmin etmek amacıyla sağlık ekonomisi alanında yaygınlıkla kullanılması önerilen regresyon teknikleri istatistiksel daraltıcı modeller için parametre optimizasyonu yapılarak uygulanacak ve performans sonuçları hem istatistiksel testlerle hem de görsel olarak karşılaştırılacaktır.

Yukarıda uygulanması planlanan işlem adımları birlikte değerlendirildiğinde, çalışmanın sağlık ekonomisi alanında kullanılan Copas testini Türkiye'de hane halkı düzeyinde bir veri setinde ilk kez uygulanacak olması, sağlık harcamasının tahmininde bir tür ayarlanmış değişken olan ödeme gücüne yer verilmiş olması, kullanılan değişkenlerin türüne en uygun dönüşüm yöntemlerinin uygulanmış olması, farklı regresyon yöntemleri arasında en iyi performans sergileyen tekniğin ve bu teknik için optimal parametre değerinin araştırılması gibi pek çok konuda sağlık harcamalarının modellemeye yönelik orijinal katkılar sunması beklenmektedir.

YÖNTEM

VERİ SETİ

Bu çalışmada veri seti olarak Türkiye İstatistik Kurumu Hanehalkı 2012 yılı Bütçe Anketi verileri kullanılmıştır.[30] Analizlerde kullanılan değişkenler; hane halkı aylık toplam cepten sağlık harcaması, hane halkı aylık toplam tüketim harcaması ve aylık toplam gıda harcaması, hane halkı reisinin cinsiyet, 65 yaş üzerinde olma durumu, sağlık sigortasına sahip olma durumu ile kır ya da kentte ikamet etme durumu olarak belirlenmiştir. Ödeme gücü değişkeninin bulunmasında Xu tarafından geliştirilen metodolojiden faydalanılmıştır.[5] Bu amaçla eş değer hane halkı büyüklüğü (*eqsize*) belirlenmiş, beta katsayısı olarak 0,56 kullanılmıştır. Eş değer hane halkı büyüklüğünün hesaplanması sonrasında yoksulluk sınırı tespit edilmiştir. Gıda için yapılan harcamaların toplam tüketim harcaması içindeki payına göre sıralı %45-55 aralığında kalan hanelerin, eş değer fert başına gıda amaçlı yapılan harcamalarının ortalaması yoksulluk sınırı olarak isimlendirilmiştir. Hanenin geçinme sınırı haricinde kalan harcama değeri, ödeme gücü olarak nitelendirilmiştir. Bu çalışmanın bağımsız değişkenlerinden birisi olarak ödeme gücü değişkeninin seçilmesinin nedeni, doğrudan gelir değişkeninin kullanılması yerine ödeme gücü değişkeninin gıda gibi zorunlu harcamaları dışarı tutan ve bir tür ayarlanmış hane halkı büyüklüğü göstergesini kullanıyor olmasıdır. Aşağıda, bu çalışmada kullanılan analiz yöntemi detaylı olarak anlatılmıştır.

ANALİZ YÖNTEMİ

Verilerin analizinde, cepten sağlık harcamalarının aylık toplam tüketim harcaması ve aylık toplam gıda harcaması değişkenleri kullanılarak hesaplanan ödeme gücü değişkeni ile hane reisinin cinsiyet, 65 yaş ve sağlık sigortası sahipliği ile hanenin kır ya da kentte ikamet etme durumu değişkenleri aracılığıyla tahmin edilmesinde regresyon modellerinden faydalanılmıştır.

Çalışmada kullanılan cepten sağlık harcaması ve ödeme gücü gibi sürekli sayısal değişkenlere ait dağılımların aşırı sağa çarpık dağılım göstermeleri nedeniyle bu değişkenlere BC dönüşümleri uygulanmış ve değişken dağılımları normal dağılıma yaklaştırılmıştır. Bu çalışmada, bağımlı değişken olarak kullanılan cepten sağlık harcaması değişkenine ait dağılımın dönüşüm sonrası normalliği, hem histogram grafiği hem de Anderson-Darling normallik testi kullanılarak incelenmiştir. Normal dağılıma uygunluk testi R programında "*nortest*" paketi kullanılarak yapılmıştır.[31] Anderson-Darling testi, özellikle ekonometrik analizlerde sıkça kullanılan ve önerilen bir normal dağılım testidir.[32] Sağlık ekonomisi alanında uzman araştırmacılar tarafından önerilen bir yaklaşım, sağlık harcamasının tahmin edildiği modellerin aşırı uyum (*overfitting*) sorunu bakımından incelenmesidir. Bunun için Copas testinin kullanıl-

ması önerilmektedir. Bu öneriye uygun olarak bu çalışmada öncelikle orijinal veri setinin %50, %55, %60, %65, %70, %75, %80, %85, %90 ve %95'ini temsil edecek şekilde 10 farklı eğitim ve test veri setleri oluşturulmuştur. R programında "*LDdiag*" paketi kullanılarak Copas testinin işlem adımları Şekil 1'e uygun şekilde izlenmiştir.[33] Copas testi sonucunda aşırı uyum sorununun en az olduğu ve en yüksek R^2 değerine sahip olan veri seti seçilmiş ve cepten sağlık harcamasının tahmin edilmesinde en küçük kareler (EKK) regresyonu ile Lasso ve Ridge gibi istatistiksel daraltıcı (Shrinkage) tahmin edicilerinin performansı, eğitim ve test veri setleri için ayrı ayrı gözlemlenmiştir. Regresyonda tahminleri dengeledikleri için "*büzülme modelleri*" olarak da isimlendirilen Lasso ve Ridge regresyonun uygulanmasında R programında "*islasso*" ve "*lmridge*" paketleri kullanılmıştır.[34,35] Regresyon modeli performanslarının karşılaştırılmasında ise MSE değerinden faydalanılmıştır. Daha sonra en iyi performansa sahip regresyon modeli kullanılarak, cepten sağlık harcamasını tahmin etmede ön plana çıkan bağımsız değişkenler, eğitim ve test veri setleri için "*ggplot2*", "*ggiraph*", "*ggiraphExtra*" paketlerinden faydalanılarak, R programında görselleştirilmiştir.[36-38] Son olarak Lasso regresyonda en iyi düzeltme parametresini (lambda-λ) bulabilmek için çapraz geçerlilikten faydalanılmış ve optimal lambda (λ) değeri eğitim ve test veri setleri için elde edilmiştir.[29]

BULGULAR

TANIMLAYICI BULGULAR

Bu çalışmada, bağımlı değişken olarak kullanılan hane halkı cepten sağlık harcaması değişkeni ile bağımsız değişken olarak kullanılan ödeme gücü değişkeni ile hane reisi ve hane halkına ait diğer bağımsız değişkenler için tanımlayıcı bilgiler Tablo 1'de verilmiştir. Bu tabloda yer verilen değişkenler arasında, ödeme gücü değişkeninin bulunmasında faydalanılan değişkenler de vardır. Buna göre, 2012 yılı için eş değer fert başına gıda harcaması ortalamasını gösteren yoksulluk sınırının 229₺ olduğu belirlenmiştir. Bunun yanı sıra 2012 yılı için ortalama hane halkı büyüklüğü 3,73 (±1,88), toplam aylık tüketim harcaması ortalama 2.545,57 (±2120,52), toplam aylık cepten sağlık harcaması ortalama 67,89 (±181,34), gıda ve alkolsüz içecek harcaması 498,15 (±353,34), aylık ödeme gücü ortalama 2348,94 (±2044,25)'tür. Kategorik formdaki çalışma değişkenleri incelendiğinde hane reisinin erkek 5.568 (%87,6) olduğu, 65 yaşın altında 5.309 (%83,5) olduğu ve sağlık sigortasına 6.027 (%94,8) sahip olduğu görülmektedir. Ayrıca hanelerin daha çok kentlerde ikamet ettikleri 4.396 (%69,1) anlaşılmaktadır.

TABLO 1: Çalışmada kullanılan değişkenlere ilişkin tanımlayıcı bilgiler.

Sayısal Değişkenler	**n**	**Minimum**	**Maksimum**	**Ortalama**	**SS**
Hane halkı büyüklüğü	6.358	1	30	3,73	1,88
Toplam tüketim harcaması	6.358	127	34.070	2.545,57	2.120,52
Gıda ve alkolsüz içecek harcaması	6.358	4,98	6.398,71	498,15	353,54
Cepten sağlık harcaması	6.358	0,05	4.513	67,89	181,34
Ödeme gücü	6.358	127,28	33.732,66	2.348,94	2.044,25
Kategorik değişkenler	**n**	**%**	**Kategorik değişkenler**	**n**	**%**
Cinsiyet			**Sigorta**		
Kadın	790	12,4	Var	6.027	94,8
Erkek	5.568	87,6	Yok	331	5,2
65 yaş ve üzerinde olma			**Kır/kent**		
65 yaş ve üzeri	1.049	16,5	Kır	1.962	30,9
65 yaş altı	5.309	83,5	Kent	4.396	69,1
Toplam	**6.358**	**100**	**Toplam**	**6.358**	**100**

EĞİTİM VE TEST VERİ SETLERİNİN OLUŞTURULMASI

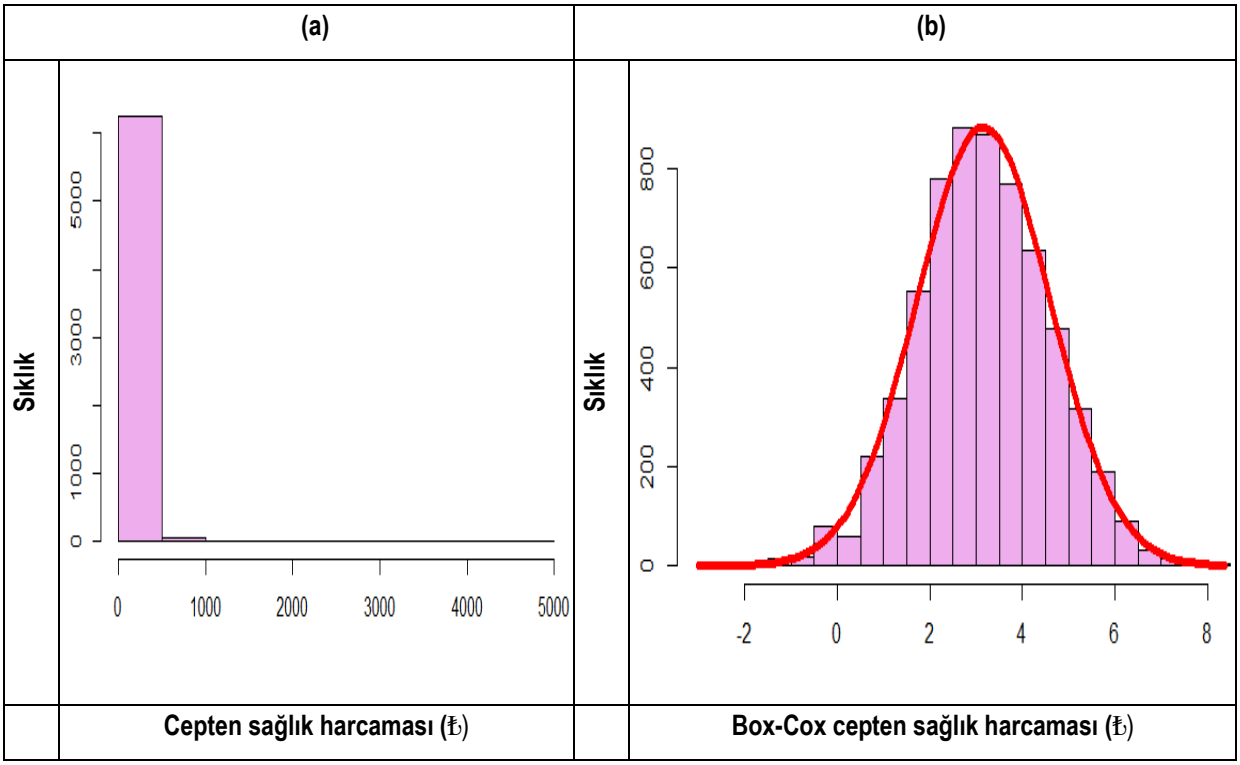
Orijinal veri setinde, cepten sağlık harcamasında bulunan toplam 6.358 hane halkı yer almaktadır. Bu çalışmada eğitim ve test veri setlerinin oluşturulması aşamasında eğitim veri seti orijinal veri setinin sırası ile %50, %55, %60, %65, %70, %75, %80, %85, %90, %95'ini temsil edecek şekilde 10 farklı eğitim (A grubu) ve test (B grubu) veri setleri oluşturulmuş ve her bir grupta bulunan gözlem sayıları kaydedilmiştir (Tablo 2).

TABLO 2: Eğitim ve test veri setlerinin oluşturulması.

Veri seti	Orijinal veri setinin %'si	Eğitim ve test veri seti	Gözlem sayıları	Veri seti	Orijinal veri setinin %'si	Eğitim ve test veri seti	Gözlem sayıları
$Veri_1$	%50	Eğitim	n_A=3.179	$Veri_6$	%75	Eğitim	n_A=4.768
	%50	Test	n_B=3.179		%25	Test	n_B=1.590
$Veri_2$	%55	Eğitim	n_A=3.497	$Veri_7$	%80	Eğitim	n_A=5.086
	%45	Test	n_B=2.861		%20	Test	n_B=1.272
$Veri_3$	%60	Eğitim	n_A=3.815	$Veri_8$	%85	Eğitim	n_A=5.404
	%40	Test	n_B=2.543		%15	Test	n_B=954
$Veri_4$	%65	Eğitim	n_A=4.133	$Veri_9$	%90	Eğitim	n_A=5.722
	%35	Test	n_B=2.225		%10	Test	n_B=636
$Veri_5$	%70	Eğitim	n_A=4.451	$Veri_{10}$	%95	Eğitim	n_A=6.040
	%30	Test	n_B=1.907		%5	Test	n_B=318

CEPTEN SAĞLIK HARCAMASI DEĞİŞKENİNE AİT DAĞILIMIN BOX-COX DÖNÜŞÜMÜ KULLANILARAK NORMALLEŞTİRİLMESİ

Cepten sağlık harcaması değişkenine ait dağılımın, normal dağılıma dönüştürülmesi amacıyla literatürde yaygınlıkla kullanılan dönüşümün, logaritmik dönüşümün bir türü olan BC dönüşümü olduğu belirtilmektedir.[16] Şekil 2 (a)'deki histogram grafiği incelendiğinde, orijinal cepten sağlık harcaması değişkenine ait dağılımın, literatürle uyumlu şekilde aşırı derece sağa çarpık özellik gösterdiği fark edilmektedir.[16] Bu nedenle cepten sağlık harcaması değişkenine ait dağılımı normalleştirmek amacıyla BC dönüşümü kullanılmıştır.[16] Şekil 2 (b)'de orijinal cepten sağlık harcaması değişkenine BC dönüşümü uygulandıktan sonraki dağılım ve yoğunluk eğrisi görülmektedir. Histogram grafikleri incelendiğinde, BC dönüşümünün dağılımı normale oldukça yaklaştırdığı görülmektedir. Bu nedenle bir sonraki aşamada, Anderson-Darling normallik testi kullanılarak, BC dönüşümü öncesi ve sonrasında orijinal veri setinden türetilen 10 farklı eğitim ve test veri seti için cepten sağlık harcaması değişkenine ait dağılımın normalliğe uygunluğu test edilmiştir.



ŞEKİL 2: Cepten sağlık harcaması değişkeninin Box-Cox dönüşümü uygulanmadan önceki ve sonraki dağılımları.

Tablo 3'te eğitim ve test veri setleri için cepten sağlık harcaması dağılımına ait Anderson-Darling normallik istatistiksel test sonuçları sunulmuştur. Eğitim veri seti orijinal veri setinin %50-95'ini temsil etmek üzere türetilen 10 farklı veri seti için BC dönüşümü öncesi ve sonrasına ait Anderson-Darling normallik test sonuçları gösterilmiştir. Normallik ölçümünde anlamlılık değeri 0,05 olarak belirlenmiştir. Anderson-Darling normallik testi sonucunda elde edilen önem düzeyinin 0,05'ten büyük olması, cepten sağlık harcamasına ait dağılımın normal dağılımdan önemli derecede farklı olmadığını göstermektedir. Buna göre, tüm veri setleri için dönüşüm öncesi cepten sağlık harcaması değişkenine ait dağılımlar normal dağılıma uygun değil iken ($p<2,2e-16$), BC dönüşümü sonrasında 1, 3, 6, 7, 8 ve 10. veri setlerinin eğitim verileri için BC dönüşümü uygulanmış cepten sağlık harcaması değişkenine ait dağılımlar normal dağılıma uygun olmayıp ($p<0,05$), test verileri için BC dönüşümü uygulanmış cepten sağlık harcaması değişkenine ait dağılımlar normal dağılıma uygundur ($p>0,05$). Bunun yanı sıra 2, 4, 5 ve 9. veri setlerinin hem eğitim ve hem de test verileri için BC dönüşümü uygulanmış cepten sağlık harcaması değişkenine ait dağılımların normal dağılımdan önemli derecede farklı olmadığı ($p>0,05$) görülmektedir.

Cepten sağlık harcaması gibi bir tür nadir olay (rare events) [yüksek harcamanın az-düşük harcamanın fazla olduğu] niteliği taşıyan değişkenler söz konusu olduğunda, regresyon sonuçlarının bağımlı değişkenin çarpıklık derecesine karşı duyarlı olduğu belirtilmektedir. Bu nedenle sağlık harcaması dağılımının yapısına en uygun dönüşümlerin uygulanması tahmin modeli performansının iyileştirilmesi için tavsiye edilen bir yaklaşımdır.[16] Bu nedenle bir sonraki aşamada regresyon modellerinin oluşturulmasında, model performanslarını iyileştirmek amacıyla cepten sağlık harcaması değişkeninin BC dönüşümü uygulanmış versiyonu kullanılmıştır.

TABLO 3: Eğitim ve test veri setleri için cepten sağlık harcaması dağılımına ait normallik testlerinin sonuçları.

Veri seti	Gözlem sayısı (Orijinal veri setinin %'si)	Box-Cox dönüşümü öncesi/ sonrası	AD*	p değeri	Veri seti	Gözlem sayısı (Orijinal veri setinin %'si)	Box-Cox dönüşümü öncesi/ sonrası	AD*	p değeri
$Veri_1$	n_A=3.179	csh	621,42	<2,2e-16	$Veri_6$	n_A=4.768	csh	900,56	<2,2e-16
	%50	BC_csh	0,8307	0,0321		%75	BC_csh	0,9034	0,0212
	n_B=3.179	csh	578,27	<2,2e-16		n_B=1.590	csh	303,16	<2,2e-16
	%50	BC_csh	0,5879	0,1254		%25	BC_csh	0,3773	0,4089
$Veri_2$	n_A=3.497	csh	683,38	<2,2e-16	$Veri_7$	n_A=5.086	csh	968,34	<2,2e-16
	%55	BC_csh	0,5468	0,1595		%80	BC_csh	0,8291	0,0324
	n_B=2.861	csh	513,7	<2,2e-16		n_B=1.272	csh	231,68	<2,2e-16
	%45	BC_csh	0,5337	0,1720		%20	BC_csh	0,3677	0,4302
$Veri_3$	n_A=3.815	csh	687,15	<2,2e-16	$Veri_8$	n_A=5.404	csh	999,51	<2,2e-16
	%60	BC_csh	0,9591	0,0155		%85	BC_csh	0,8641	0,0266
	n_B=2.543	csh	508,3	<2,2e-16		n_B=954	csh	197,83	<2,2e-16
	%40	BC_csh	0,2807	0,6419		%15	BC_csh	0,2561	0,7237
$Veri_4$	n_A=4.133	csh	816,28	<2,2e-16	$Veri_9$	n_A=5.722	csh	1060,6	<2,2e-16
	%65	BC_csh	0,5365	0,1693		%90	BC_csh	0,7314	0,0565
	n_B=2.225	csh	365,62	<2,2e-16		n_B=636	csh	133,77	<2,2e-16
	%35	BC_csh	0,5746	0,1356		%10	BC_csh	0,3474	0,4782
$Veri_5$	n_A=4.451	csh	836,16	<2,2e-16	$Veri_{10}$	n_A=6.040	csh	1.141	<2,2e-16
	%70	BC_csh	0,8344	0,0314		%95	BC_csh	0,8539	0,0282
	n_B=1.907	csh	364,67	<2,2e-16		n_B=318	csh	59,957	<2,2e-16
	%30	BC_csh	0,2727	0,6683		%5	BC_csh	0,3603	0,4458

Açıklamalar: n_A: Eğitim veri setindeki gözlem sayısı; n_B: Test veri setindeki gözlem sayısı; csh: dönüşüm öncesi cepten sağlık harcaması değişkeni; BC_csh: Box-Cox dönüşümü uygulandıktan sonra cepten sağlık harcaması değişkeni; *AD: Anderson-Darling normallik testi.

ANALİTİK BULGULAR

Bu çalışmada, cepten sağlık harcaması değişkenini tahmin etmek amacıyla regresyon modelinin oluşturulmasından önce değişkenler arası ilişki çoklu bağlantı bakımından Spearman korelasyon katsayısı (r_s) kullanılarak incelenmiştir. Bu çalışmanın bağımsız değişkenleri arasında yer alan ödeme gücü, cinsiyet, 65 yaş üzerinde olma durumu, sağlık sigortasının varlığı, kır ya da kentte ikamete etme durumu ve bağımlı değişken olan cepten sağlık harcaması değişkenleri arasındaki ilişki katsayılarının 0,75'ten küçük olduğu görülmektedir. Bu nedenle çoklu bağlantı sorununun olmadığı ve tüm bağımsız değişkenlerin modele dâhil edilebileceği sonucuna varılmıştır. Bu çalışmada, cepten sağlık harcaması değişkenini tahmin etmek amacıyla oluşturulan regresyon modeli Eşitlik 10'da gösterilmiştir. Buna göre BC dönüşümü uygulanmış cepten sağlık harcaması değişkenini tahmin etmek amacıyla 5 farklı bağımsız değişkenden oluşan bir model oluşturulmuştur. Bu model oluşturulurken, ödeme gücü değişkeni için de BC dönüşümü uygulanmıştır. Copas testinin aşamalarından birisi regresyon modelinin öncelikle eğitim veri seti üzerinde uygulanmasıdır. Bu aşamadan sonra eğitim veri seti üzerinde elde edilen tahmin sonuçları kullanılarak, aynı tahmin modeli test veri seti üzerinde çalıştırılmaktadır. Bu tahmine, Eşitlik 11'de görüldüğü üzere "*Yhat*" ismi verilmektedir. Bir sonraki aşamada ise Eşitlik 12'de görüldüğü üzere, test veri seti üzerinde BC dönüşümü uygulanmış cepten sağlık harcaması değişkeninin bağımlı, "*Yhat*" değişkeninin ise bağımsız değişken olduğu bir regresyon modeli oluşturulmaktadır. Oluşturulan bu model sonucunda farksızlık hipotezi (β=1) kabul edildiğinde, aşırı uyum (*overfitting*) so-

rununun olmadığı kararı verilmektedir. Aşırı uyum sorununun üstesinden gelebilmek için veri setinin, farklı gruplara bölünmesi ve farklı türde regresyon modellerinin denenmesi önerilmektedir.[19,23,24]

$$fit.lm=[Box-Cox(csh)_{it} = \alpha_{it} + \beta_1 Box-Cox(odgucu)_{it} + \beta_2(cinsiyet)_{it} + \beta_3(65\ yaş)_{it} + \beta_4(sigorta)_{it} + \beta_5(Kır/Kent)_{it} + \varepsilon_{it}\ ] \quad \textbf{(10)}$$

$$Yhat = predict\ (fit.lm, test\ veri\ seti) \quad \textbf{(11)}$$

$$newfit = lm\ (Box-Cox_csh)_{test\ veri\ seti} \sim Yhat \quad \textbf{(12)}$$

COPAS TESTİ SONUÇLARI

Cepten sağlık harcaması değişkenini tahmin etmek amacıyla BC dönüşümü uygulanmış cepten sağlık harcaması değişkeninin, tahmin edici değişken olarak yer aldığı ve eğitim veri setleri üzerinde oluşturulup, test veri setleri üzerinde sınanan Copas testi sonucunda 10 farklı test veri setinden elde edilen regresyon bulguları Tablo 4'te sunulmuştur. Buna göre 1, 9 ve 10. veri setlerinden elde edilen Copas testleri sonucunda β değeri 1 olarak bulunmuş ve aşırı uyumun olmadığı görülmüştür. Bunun yanı sıra 9. veri setinden elde edilen R^2 ve düzeltilmiş R^2 değerleri en yüksek olup, daha iyi model performansı sergilemiştir. Bu doğrultuda bir sonraki aşamada eğitim veri seti orijinal veri setinin %90'ını temsil etmek üzere cepten sağlık harcamasının tahmin edici değişken olduğu, çoklu regresyon modelleri 9. veri seti üzerinde uygulanmış ve model performansları karşılaştırılmıştır.

TABLO 4: Copas testi sonucunda 10 farklı test veri setinden elde edilen regresyon bulguları.

Veri seti[&]	Tahmin edilen değişken[&]	Beta (β)	t	*se*	*p*	R^2	*Düzeltilmiş* R^2	*F*	*p değeri*
VS_1-test verisi	*yhat*[*]	1,17515	18,57	0,06329	<0,0001	0,0979	0,09761	344,8	<0,0001
VS_2-test verisi	*yhat*[*]	0,87603	16,728	0,05237	<0,0001	0,08915	0,08883	279,8	<0,0001
VS_3-test verisi	*yhat*[*]	0,78848	16,160	0,04879	<0,0001	0,0932	0,09284	261,2	<0,0001
VS_4-test verisi	*yhat*[*]	0,65854	14,992	0,04393	<0,0001	0,09182	0,09141	224,8	<0,0001
VS_5-test verisi	*yhat*[*]	0,73968	13,576	0,05449	<0,0001	0,08821	0,08773	184,3	<0,0001
VS_6-test verisi	*yhat*[*]	0,89947	13,351	0,06737	<0,0001	0,10090	0,10030	178,2	<0,0001
VS_7-test verisi	*yhat*[*]	0,82211	11,068	0,07428	<0,0001	0,08796	0,08725	122,5	<0,0001
VS_8-test verisi	*yhat*[*]	0,90123	9,853	0,09147	<0,0001	0,09253	0,09158	97,07	<0,0001
VS_9-test verisi	*yhat*[*]	1,09850	9,542	0,1151	<0,0001	0,12560	0,12420	91,05	<0,0001
VS_{10}-test verisi	*yhat*[*]	1,60420	5,838	0,2748	<0,0001	0,09735	0,09449	34,08	<0,0001

$*: Yhat = predict\ (fit.lm, test\ veri\ seti)$; [&]:Copas testi işlem adımları, tahmin değişkeni, veri setleri ve regresyon modelleri hakkında ayrıntılı bilgi için Şekil 1, Tablo 2 ve Eşitlik 10, 11 ve 12'yi inceleyiniz.

REGRESYON SONUÇLARI

Eşitlik 13, 14 ve 15'te EKK, Lasso ve Ridge regresyon modelleri gösterilmiştir.

EKK regresyonu için model:

$$fit.lm=[Box-Cox(csh)_{it}= \alpha_{it}+\beta_1 Box-Cox(odgucu)_{it}+\beta_2(cinsiyet)_{it}+\beta_3(65\ yaş)_{it}+\beta_4(sigorta)_{it}+\beta_5(Kır/Kent)_{it}+\varepsilon_{it}\ ] \quad \textbf{(13)}$$

Lasso regresyon için model:

$$fit.lasso=[Box-Cox(csh)_{it}= \alpha_{it}+\beta_1 Box-Cox(odgucu)_{it}+\beta_2(cinsiyet)_{it}+\beta_3(65\ yaş)_{it}+\beta_4(sigorta)_{it}+\beta_5(Kır/Kent)_{it}+\varepsilon_{it}\ ] \quad \textbf{(14)}$$

Ridge regresyon için model:

$$fit.ridge=[Box-Cox(csh)_{it}= \alpha_{it}+\beta_1 Box-Cox(odgucu)_{it}+\beta_2(cinsiyet)_{it}+\beta_3(65\ yaş)_{it}+\beta_4(sigorta)_{it}+\beta_5(Kır/Kent)_{it}+\varepsilon_{it}\ ] \quad \textbf{(15)}$$

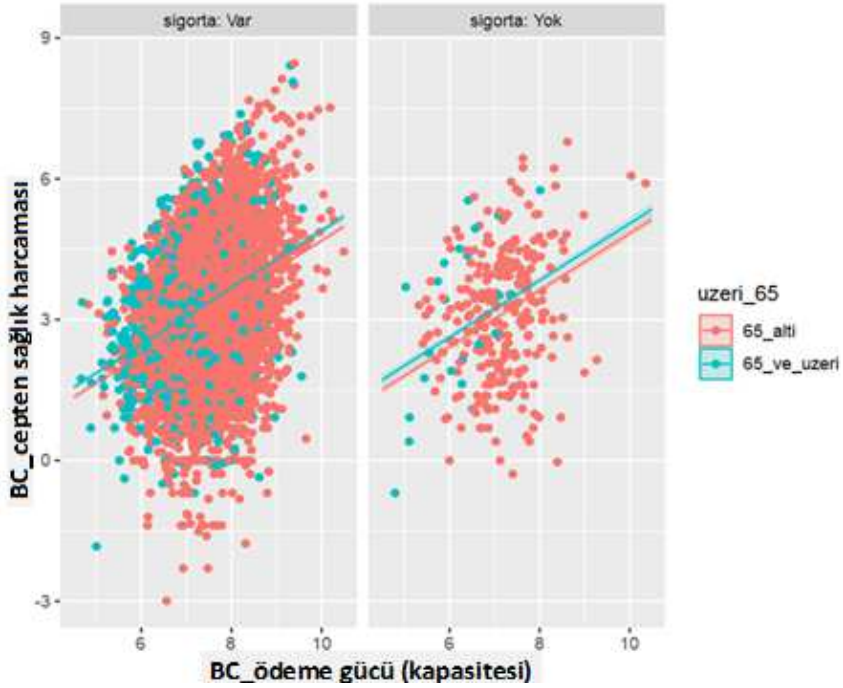
Copas testinde iyi performans sergilediği görülen 9 no.lu veri seti için eğitim (n_A=5.722) ve test (n_B=636) veri setleri üzerinde cepten sağlık harcamasını tahmin etmek amacıyla oluşturulan regresyon modellerine ait tahmin sonuçları ve MSE değerleri Tablo 5'te sunulmuştur. Buna göre hem eğitim hem de test veri seti üzerinde yapılan uygulamalarda EKK regresyonu en düşük MSE değerlerini vermiştir. Büzülme modelleri olarak isimlendirilen, Lasso ve Ridge regresyon karşılaştırıldığında ise Lasso regresyonun Ridge tahmin ediciye kıyasla tahminleri büzerek daha küçük MSE değerleri ve daha iyi tahmin performansı ortaya koyduğu görülmektedir. EKK modeli kullanılarak eğitim ve test veri setleri için en yüksek ilk 3 tahmin edici değişken arasındaki ilişkinin görsel sunumu Şekil 3 ve 4'te gösterilmiştir. Bir sonraki aşamada ise Ridge regresyon yöntemine göre daha iyi tahmin performansı sergilediği görülen Lasso regresyonda en iyi büzülme parametresi değerine ulaşmak için çapraz geçerlilikten faydalanılmış ve sonuçlar Şekil 5 ve 6'da sunulmuştur.

TABLO 5. Dokuz no.lu veri seti için eğitim ve test veri setleri kullanıldığında model performanslarının karşılaştırması.

Beta (β)	Eğitim veri seti (n_A=5.722)			Test veri seti (n_B=636)		
	EKK	Lasso	Ridge	EKK	Lasso	Ridge
β_1 *(ödeme gücü)*	**0,6094**	**0,5884**	0,6040	**0,6454**	**0,5480**	0,6454
β_2 *(cinsiyet)*	0,0173	0,0055	0,0162	**0,0599**	**0,0130**	0,0600
β_3 *(65 yaş üzeri)*	**0,2075**	**0,1707**	0,2038	**0,2068**	**0,0250**	0,2068
β_4 *(sigorta)*	**0,1537**	**0,0955**	0,1506	0,0192	-0,0063	0,0192
β_5 *(kır-kent)*	0,0145	0,0030	0,0121	-0,0923	-0,0202	-0,0924
Hata kareler ortalaması- MSE	**1,8827**	**1,8830**	10,3906	**1,5900**	**1,6000**	8,8488

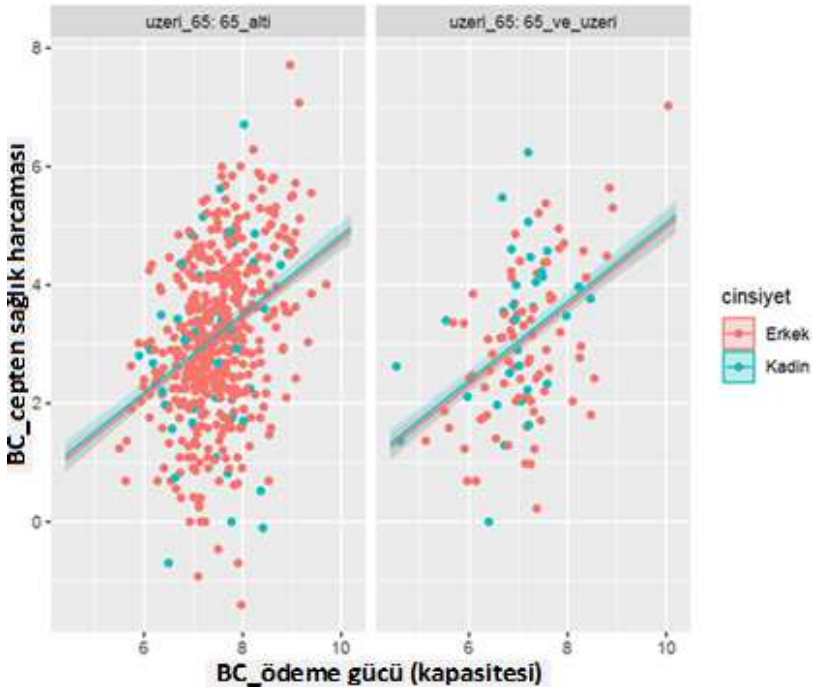
EKK: En küçük kareler; MSE: Mean squared error.

Şekil 3'te eğitim veri seti kullanılarak oluşturulan EKK modelinde, cepten sağlık harcamasını (BC_cepten sağlık harcaması) tahmin etmek üzere, ödeme gücü (BC_ödeme gücü), hane reisinin 65 yaş üzeri olma ve sağlık sigortasına sahip olma durumundan oluşmak üzere 3 tahmin edici değişkenin etkileşimi görülmektedir. Buna göre incelenen hane halkları içerisinde, hane reisi sağlık sigortasına sahip olanlar ağırlıklı olmakla birlikte, cepten sağlık harcaması yapanları daha çok 65 yaş ve üzerinde olan grup temsil etmektedir.



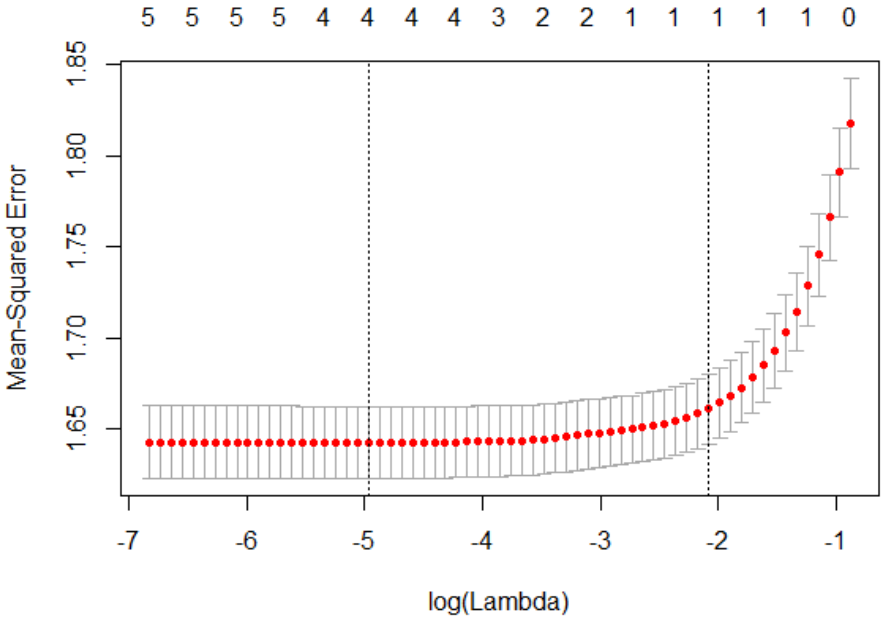
ŞEKİL 3: EKK regresyonda eğitim veri seti kullanılarak cepten sağlık harcamasının tahmini.

Şekil 4'te test veri seti kullanılarak oluşturulan EKK modelinde, cepten sağlık harcamasını (BC_cepten sağlık harcaması) tahmin etmek üzere, ödeme gücü (BC_ödeme gücü), cinsiyet ve 65 yaş üzeri olma durumundan oluşmak üzere 3 tahmin edici değişkenin etkileşimi görülmektedir. Buna göre incelenen hane halkları içerisinde hane reisinin daha çok 65 yaş altı grupta bulunduğu görülmekle birlikte, cepten sağlık harcaması yüksek olanlar daha çok kadın cinsiyete sahip hane reislerinin olduğu hanelerdir.



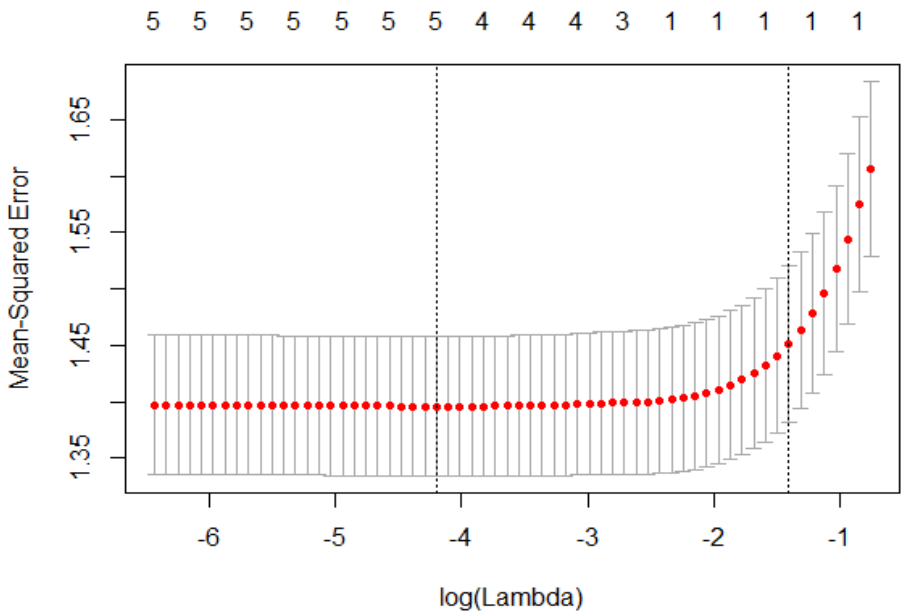
ŞEKİL 4: EKK regresyonda test veri seti kullanılarak cepten sağlık harcamasının tahmini.

Bu çalışmada EKK'den sonra en iyi tahmin gücüne sahip olan istatistiksel daraltıcı (Shrinkage) modelin Lasso regresyon olduğu görülmüştür. Lasso regresyonda, en iyi büzülme değerinin elde edilmesi sayesinde MSE en aza indirilebilmekte, bu amaçla çapraz geçerlilikten faydalanılabilmektedir. Lasso regresyonda optimal λ değerinin bulunması için daha önce Copas testinden iyi performans sergilediği belirlenen ve eğitim verisinin orijinal veri setinin %90'ını temsil ettiği 9 no.lu veri kullanılmıştır. Eğitim veri seti için cepten sağlık harcamasını tahmin etmek amacıyla oluşturulan Lasso regresyon modelinde, çapraz geçerlilik uygulandığında en iyi λ değeri 0,0158 olarak bulunmuştur. En iyi λ değeri kullanılarak yapılan tahminde, cepten sağlık harcamasını tahmin etmede en etkili değişkenin ise ödeme gücü (β_1=0,5046) olduğu görülmüştür. Şekil 5'te 9 no.lu verinin eğitim veri seti kullanıldığında Log (λ)'nın bir fonksiyonu olarak, Lasso regresyon için MSE'nin çapraz geçerlilik ile tahmini gösterilmiştir. Bu şekil üzerinde yatay eksen log (λ) değerine, dikey eksen ise MSE değerine işaret etmektedir. Şekil 5 üzerinde her bir kırmızı noktaya ait gri alanlar MSE_λ için eksi ve artı bir standart hata değerlerine işaret eder. Dikey kesikli çizgilerden en soldaki ise minimum MSE değerini temsil etmektedir.



ŞEKİL 5: Dokuz no.lu veride eğitim veri seti kullanıldığında log (λ)'nın bir fonksiyonu olarak Lasso regresyon için MSE'nin çapraz geçerlilik ile tahmini.

Lasso regresyonda, optimal λ değerinin bulunması için 9 no.lu veri setinin test verisi (orijinal verinin %10'u) kullanılarak cepten sağlık harcamasını tahmin etmek amacıyla oluşturulan Lasso regresyon modelinde, çapraz geçerlilik uygulandığında en iyi λ değerinin 0,0630 olduğu tespit edilmiştir. En iyi λ değeri kullanılarak yapılan tahminde, cepten sağlık harcamasını tahmin etmede en etkili değişken ise ödeme gücü (β_1=0,5579) olarak bulunmuştur. Şekil 6'da 9 no.lu veri setinin test verisi kullanıldığında log (λ)'nın bir fonksiyonu olarak Lasso regresyon için MSE'nin çapraz geçerlilik ile tahmini gösterilmiştir. Bu şekil üzerinde yatay eksen log (λ) değerlerini, dikey eksen ise MSE değerlerini göstermektedir. Şekil 6 üzerinde her bir kırmızı noktaya ait gri alanlar MSE_λ için eksi ve artı bir standart hata değerlerini ifade etmektedir. Dikey kesikli çizgilerden en soldaki ise minimum MSE değerine karşılık gelmektedir.



ŞEKİL 6: Dokuz no.lu veride test veri seti kullanıldığında log (λ)'nın bir fonksiyonu olarak Lasso regresyon için MSE'nin çapraz geçerlilik ile tahmini.

TARTIŞMA

Regresyon modellerinde bağımlı ve bağımsız değişken arasında aşırı uyumun bulunması, regresyon modelinin genel performansını düşürücü bir etkide bulunmakta ve modele ilişkin yapılacak olan çıkarsama ve öngörüler açısından sorun yaratmaktadır.[10,39] Böylesi bir durumda model genelleme yeteneğini kaybetmektedir. Bu nedenle model performansını yükseltmeye yönelik çeşitli tekniklerden faydalanılmaktadır.[40] Bu tekniklerden birisi olarak sağlık ekonomistleri tarafından önerilen bir yaklaşım, Copas testinin uygulanmasıdır.[10] Ayrıca Lasso ve Ridge regresyon gibi tahminleri dengeleme özelliğine sahip olan büzülme modelleri de kullanılabilmektedir.[13,14]

Sağlık harcaması ya da sağlık ile ilgili maliyetlerin incelendiği çalışmalarda, regresyon modellerinin oluşturulmasında sıklıkla karşılaşılan problemlerin başında dağılımın normal dağılıma uygun olmaması gelmektedir.[41] Bu nedenle farklı dönüşüm yöntemleri uygulanarak, normal dağılıma uygunluk sağlanmaktadır. Bu sayede oluşturulacak olan regresyon modellerinin performansının yükseltilmesi ve sağlık harcamalarını belirlemeye yönelik daha güçlü tahminlerin yapılması mümkün olmaktadır.[42]

Bu çalışmada, uygun dönüşümler uygulandıktan sonra cepten sağlık harcamasını tahmin etmek amacıyla orijinal veri setini farklı büyüklüklerde temsil etmek üzere eğitim ve test veri setleri oluşturularak, Copas testi işlem adımları uygulanmıştır. Aşırı uyum sorununun en az olduğu eğitim ve test veri setleri seçilerek, cepten sağlık harcamasını tahmin etmek üzere eğitim ve test veri setleri için EKK, Lasso ve Ridge regresyon modelleri uygulanmıştır. Tahmin modellerinin performansları MSE ile karşılaştırılmıştır. Eğitim ve test veri setleri için EKK modelinin, daha düşük hata değerlerine ve daha iyi tahmin performansına sahip olduğu görülmüştür. Eğitim ve test veri setleri için ödeme gücü değişkeni, cepten sağlık harcamasını tahmin etmede en güçlü değişkendir (β>0,50). Farklı regresyon modelleri karşılaştırıldığında, EKK modeli en iyi tahmini gerçekleştirmekle birlikte büzülme modelleri içerisinde Lasso regresyon Ridge regresyondan daha iyi performans sergilemiştir. Son olarak çapraz geçerlilik sonucunda eğitim ve test veri setleri için Lasso regresyonda en iyi λ parametre değerleri elde edilmiştir.

Regresyon modellerinin gücünü belirlemeye yönelik yapılan araştırmalarda, genellikle veri setindeki değişken sayısının en az 3 ya da 4 katı kadar gözlemin bulundurulması gerektiği önerilmesine rağmen bu

konuda genel kabul gören bir yaklaşım yoktur.[11] Bu nedenle veri setinde yer alan değişkenlerin tanımlayıcı özellikleri ve değişkenler arası ilişkilerden yola çıkılarak birtakım optimizasyon yöntemleri uygulanmaktadır.[43] Bu çalışmada ise sağlık ekonomisi alanında, aşırı çarpıklık özelliği nedeniyle modellenmesi oldukça zor olan sağlık harcaması değişkenini tahmin etmek amacıyla bir çoklu regresyon modeli oluşturulmuştur. Daha iyi tahmin performansına erişmek için Copas testi sonucunda aşırı uyum sorunun en az olduğu eğitim ve test veri setleri seçilmiş ve regresyon modellerinin performans sonuçları karşılaştırılmıştır. EKK, Lasso ve Ridge regresyon sonuçları incelendiğinde, EKK model performansının en iyi olduğu, Lasso regresyonun ise 2. en iyi tahmin performansına sahip olduğu görülmüştür. İlerleyen çalışmalar için genelleştirilmiş regresyon modelleri ya da alternatif büzülme modellerinin de araştırmaya dâhil edilmesi yolu ile tahmin sonuçlarının karşılaştırması tavsiye edilmektedir. Ayrıca bu çalışmada, Ridge regresyona göre daha iyi tahmin performansı sergilediği gözlenen Lasso regresyonda yalnızca tek alfa (α=1) parametresi kullanılmıştır. İlerleyen araştırmalar için farklı α ve λ parametre değerleri belirlendiğinde elde edilecek tahmin sonuçları karşılaştırılabilir. Son olarak Monte Carlo simülasyonu gibi stokastik tekniklerin kullanımı ile tahmin sonuçlarının karşılaştırılmalı olarak gözden geçirilmesi de tavsiye edilebilir.

SONUÇ

Bu çalışmada, cepten sağlık harcaması değişkenini ödeme gücü, cinsiyet, 65 yaş üzerinde olma, sağlık sigortasına sahip olma ve kır ya da kentte ikamet etmek değişkenleri kullanılarak tahmin etmeye yönelik bir çoklu regresyon modeli oluşturulmuştur. Orijinal veri seti, farklı büyüklüklerde eğitim ve test veri setlerine ayrılarak Copas testi uygulanmıştır. Cepten sağlık harcamasını tahmin etmek amacıyla oluşturulan EKK, Lasso ve Ridge modelleri karşılaştırıldığında, EKK modelinin en iyi olduğu görülmüştür. Lasso regresyon ise Ridge regresyondan daha iyi tahmin performansı sergilemiştir. İlerleyen araştırmalar için farklı regresyon yöntemleri, farklı büyüklük ve türdeki veri setleri ile parametre optimizasyon teknikleri kullanılarak sonuçların karşılaştırılması önerilmektedir.

Finansal Kaynak

Bu çalışma sırasında, yapılan araştırma konusu ile ilgili doğrudan bağlantısı bulunan herhangi bir ilaç firmasından, tıbbi alet, gereç ve malzeme sağlayan ve/veya üreten bir firma veya herhangi bir ticari firmadan, çalışmanın değerlendirme sürecinde, çalışma ile ilgili verilecek kararı olumsuz etkileyebilecek maddi ve/veya manevi herhangi bir destek alınmamıştır.

Çıkar Çatışması

Bu çalışma ile ilgili olarak yazarların ve/veya aile bireylerinin çıkar çatışması potansiyeli olabilecek bilimsel ve tıbbi komite üyeliği veya üyeleri ile ilişkisi, danışmanlık, bilirkişilik, herhangi bir firmada çalışma durumu, hissedarlık ve benzer durumları yoktur.

Yazar Katkıları

Bu çalışma tamamen yazarın kendi eseri olup başka hiçbir yazar katkısı alınmamıştır.

REFERENCES

1. Grigoli F, Kapsoli J. Waste not, want not: The efficiency of health expenditure in emerging and developing economies. Rev Dev Econ. 2018;22(1):384-403.[Crossref] [PubMed] [PMC]
2. Newhouse JP. Medical-care expenditure: a cross-national survey. J Hum Resour. 1977;12(1):115-25.[Crossref] [PubMed]
3. Catlin MK, Poisal JA, Cowan CA. Out-of-pocket health care expenditures, by insurance status, 2007-10. Health Aff (Millwood). 2015;34(1):111-6.[Crossref] [PubMed]
4. Xu K, Evans DB, Kawabata K, Zeramdini R, Klavus J, Murray CJ. Household catastrophic health expenditure: a multicountry analysis. Lancet. 2003;362(9378):111-7.[Crossref] [PubMed]
5. World Health Organization. Distribution of health payments and catastrophic expenditures Methodology/by Ke Xu. World Health Organization. 2005.[Link]
6. Eigner I, Hamper A. Predictive analytics in health care: methods and approaches to identify the risk of readmission. In: Wickramasinghe N, Schaffer JL, eds. Theories to Inform Superior Health Informatics Research and Practice. Healthcare Delivery in the Information Age. Cham, Switzerland: Springer; 2018. p.55-73.[Crossref]
7. Manning WG. The logged dependent variable, heteroscedasticity, and the retransformation problem. J Health Econ. 1998;17(3):283-95.[Crossref] [PubMed]

8. Manning WG, Mullahy J. Estimating log models: to transform or not to transform? J Health Econ. 2001;20(4):461-94.[Crossref] [PubMed]
9. Harrell FE. Regression Modeling Strategies with Applications to Linear Models, Logistic Regression and Survival Analysis. Springer Series in Statistics. New York: Springer-Verlag; 2001. p.103-26.[Crossref]
10. Bilger M, Manning WG. Measuring overfitting in nonlinear models: a new method and an application to health expenditures. Health Econ. 2015;24(1):75-85.[Crossref] [PubMed]
11. Alpar R. Uygulamalı Çok Değişkenli İstatistiksel Yöntemler. 3. Baskı. Ankara: Detay Yayıncılık; 2011. p.415-628.
12. Rodrigues JFD. A bayesian approach to the balancing of statistical economic data. Entropy. 2014;16(3):1243-71.[Crossref]
13. Tibshirani R. Regression shrinkage and selection via the lasso. J R Statist Soc: Series B: Methodological. 1996;58(1):267-88.[Crossref]
14. Varmuza K, Filzmoser P. Introduction to Multivariate Statistical Analysis in Chemometrics, Taylor & Francis Group, CRC Press, 2009.
15. Davison AC, Hinkley DV. Further topics in regression. In: Bootstrap Methods and their Application. Cambridge Series in Statistical and Probabilistic Mathematics. New York, USA: Cambridge University Press; 1997. p.326-84.[Crossref]
16. Manning W. Dealing with skewed data on costs and expenditures. In: Jones AM, ed. The Elgar Companion to Health Economics. Cheltenham, U.K.; Northampton, Mass.: Edward Elgar; 2006. p.439-46.
17. Frank IE, Friedman JH. A statistical view of some chemometrics regression tools. Technometrics. 1993;35(2):109-35.[Crossref]
18. Larson SC. The shrinkage of the coefficient of multiple correlation. J Educ Psychol. 1931;22(1):45-55.[Crossref]
19. Copas JB. Cross validation shrinkage of regression predictors. J R Statist Soc: Series B: Methodological. 1987;49(2):175-83.[Crossref]
20. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(4):361-87.[Crossref] [PubMed]
21. Harrell FE. Regression Modelling Strategies: with Applications to Linear Models, Logistic Regression and Survival Analysis. New York: Springer; 2001.[Crossref]
22. Blough DK, Madden CW, Hornbrook MC. Modeling risk using generalized linear models. J Health Econ. 1999;18(2):153-71.[Crossref] [PubMed]
23. Ellis RP, Mookim PG. Mookim. Cross-Validation Methods for Risk Adjustment Models. 2009.[Link]
24. Copas JB. Regression, prediction and shrinkage. J R Statist Soc: Series B (Methodological). 1983;45(3):311-54.[Crossref]
25. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning with Applications in R. Springer Texts in Statistics. New York: Springer: Springer Science+Business Media; 2013. p.214-27.[Crossref]
26. Hastie T, Tibshirani R, Wainwright M. Optimization methods. In: Statistical Learning with Sparsity: The Lasso and Generalizations. Chapman & Hall; 2015. p.93-127.[Crossref]
27. Friedman J, Hastie T, Tibshirani R, Narasimhan B, Tay K, Simon N, et al. Package "glmnet". R Foundation for Statistical Computing; 2020.[Link]
28. Hastie T, Tibshirani R, Friedman JH. Linear methods for regression. In: The Elements of Statistical Learning. New York: Springer; 2009. p.43-93.[Crossref]
29. Melkumova LE, Shatskikh SY. Comparing ridge and LASSO estimators for data analysis. Procedia Engineering. 2017;201:746-55.[Crossref]
30. Türkiye İstatistik Kurumu (TÜİK). Hanehalkı Bütçe Anketi-2012. (Erişim tarihi: 14.9.2020).[Link]
31. Gross J, Ligges U. Package "nortest". R Foundation for Statistical Computing; 2015.[Link]
32. Dufour JM, Farhat A, Gardiol L, Khalaf L. Simulation-based finite sample normality tests in linear regressions. Economet J. 1998;1(1):C154-73.[Crossref]
33. Yongmei N. LDdiag: link function and distribution diagnostic test for social science researchers. (2012).[Link]
34. Sottile G, Cilluffo G, Muggeo VMR. Package "islasso". R Foundation for Statistical Computing; 2020.[Link]
35. Muhammad IU, Muhammad A. Package "lmridge". R Foundation for Statistical Computing; 2018.[Link]
36. Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C, et al. Package "ggplot2". R Foundation for Statistical Computing; 2020.[Link]
37. Gohel D, Skintzos P, Bostock M, Kokenes S, Shull E, Book E. Package "ggiraph". R Foundation for Statistical Computing; 2020.[Link]
38. Moon KW. Package "ggiraphExtra". R Foundation for Statistical Computing; 2018.[Link]
39. Pacifico A. Robust open Bayesian analysis: Overfitting, model uncertainty, and endogeneity issues in multiple regression models. Econom. Rev. 2020;[Crossref]
40. Afifi AA, Kotlerman JB, Ettner SL, Cowan M. Methods for improving regression analysis for skewed continuous or counted responses. Annu Rev Public Health. 2007;28:95-111.[Crossref] [PubMed]
41. Briggs A, Gray A. The distribution of health care costs and their statistical analysis for economic evaluation. J Health Serv Res Policy. 1998;3(4):233-45.[Crossref] [PubMed]
42. Gerdtham UG, Søgaard J, Andersson F, Jönsson B. An econometric analysis of health care expenditure: a cross-section study of the OECD countries. J Health Econ. 1992;11(1):63-84.[Crossref] [PubMed]
43. Raikou M, Mcguire A. Estimating costs for economic evaluation. In: Jones AM, ed. The Elgar Companion to Health Economics. Cheltenham, U.K.; Northampton, Mass.: Edward Elgar; 2006. p.429-38.