Ekoist: Journal of Econometrics and Statistics, 33, 13-31

DOI: 10.26650/ekoist.2020.33.843564
http://ekoist.istanbul.edu.tr

Başvuru: 08.12.2020
Kabul: 30.12.2020

İSTANBUL ÜNİVERSİTESİ YAYINEVİ

Ekoist: Journal of Econometrics and Statistics

ARAŞTIRMA MAKALESİ / RESEARCH ARTICLE

Hanehalkı Tüketim Harcamalarının Mikroekonometrik Analizi: LAD-LASSO Yöntemi*

Microeconometric Analysis of Household Consumption Expenditures: LAD-LASSO Method

Kadriye Hilal Topal** , **Ebru Çağlayan Akay*****

Öz

Bu çalışmanın amacı, denetimli makine öğrenmesi yöntemlerinin aşırı değer ve uzun kuyruklu hatalara sahip Hanehalkı Bütçe Anketi Hane veri setinin ilgili değişkenlerini seçmemize nasıl yardımcı olduğunu incelemek ve Türkiye'nin Hanehalkı Tüketim Harcamaları'nın tahmininde en iyi tahmin ve öngörü performansına sahip olan modelin belirlenmesini sağlamaktır. Bu amaçla, 2018 yılı Türkiye'nin Hanehalkı Bütçe Anketi Hane veri seti klasik regresyon yönteminin yanı sıra En Küçük Mutlak Sapma (LAD), En Küçük Mutlak Küçültme ve Seçim Operatörü (LASSO) ve LAD-LASSO yöntemleri kullanılarak incelenmiş ve yöntemlerin tahmin ve öngörü performansları karşılaştırılmıştır. Analiz sonuçlarına göre; uzun kuyruklu hataların varlığında dayanıklı tahminciler elde edilirken aynı zamanda değişken seçimine olanak sağlayan LAD-LASSO makine öğrenmesi yönteminin tahmin performansı ve öngörü açıklığı açısından en başarılı yöntem olduğu sonucuna ulaşılmıştır. Ayrıca gelir, tasarruf ve hane halkı büyüklüğü gibi bazı temel değişkenler tüm modeller için hanehalkı tüketim harcamalarını artırmaktadır. Bu değişkenlere ek olarak odanın yapısı, mutfak, banyo zeminleri, ısıtma, klima tercihleri, kullanılan enerji kaynakları, müstakil ev, apartman, yazlık, bağ sahipliği ve yatırım tercihleri, kredi kartı kullanımı, internet alışveriş alışkanlıkları gibi çeşitli değişkenler LAD-LASSO modelinde hane halkı tüketim harcamalarının belirleyicileri olarak seçilmiştir. Çalışma sonuçlarından, makine öğrenme algoritmalarının mikroekonometrik modellerin oluşturulması sırasında gerekli değişkenlerin seçiminde kullanılabileceğine dair bulgular elde edilmiştir. Bu çalışma doktora tezinden üretilmiştir.

Anahtar Kelimeler
Makine Öğrenmesi, LAD-LASSO Regresyonu, Hanehalkı Tüketim Harcamaları

Jel Sınıflandırması
C31, C55, D12

Abstract

This study examined how supervised machine learning methods help us select the relevant variables of a Household Budget Survey Consumption Expenditures dataset with outliers in order to achieve better performance in the predicting and forecasting of the Household Consumption Expenditures Model. To achieve this, the Household Budget Survey Consumption Expenditures dataset of Turkey for 2018 was examined using the Least Absolute Deviation (LAD), Least Absolute Shrinkage and Selection Operator (LASSO) and LAD-LASSO methods. In addition, the classical regression method and the prediction and forecasting performances of the methods were compared. According to the analyzed results,

* Bu çalışma Kadriye Hilal Topal'ın hazırladığı doktora tezinden üretilmiştir.

** **Sorumlu Yazar:** Kadriye Hilal Topal (Öğr. Gör.), Nişantaşı Üniversitesi, Meslek Yüksekokulu, Bilgisayar Programcılığı Bölümü, İstanbul, Türkiye. E-posta: hilal.topal@nisantasi.edu.tr ORCID: 0000-0001-5203-8017

*** Ebru Çağlayan Akay (Prof. Dr.), Marmara Üniversitesi, İktisat Fakültesi, Ekonometri Bölümü, İstanbul, Türkiye. E-posta: ecaglayan@marmara.edu.tr ORCID: 0000-0002-9998-5334

Atıf: Topal, K. H. ve Caglayan-Akay, E. (2020). Hanehalkı tüketim harcamalarının mikroekonometrik analizi: LAD-LASSO yöntemi. *EKOIST Journal of Econometrics and Statistics*, 33, 13-31. https://doi.org/10.26650/ekoist.2020.33.843564

it was concluded that the LAD-LASSO machine learning method, which enables the selection of variables while obtaining robust predictors in the presence of long-tailed errors, was the most successful method in prediction performance and forecasting accuracy. Additionally, several fundamental variables such as income, saving, and household size increase the household consumption expenditures for all models. In addition to these variables, other variables including the structure of a room, the kitchen, bathroom floors, heating, air conditioning preferences, energy sources used, detached house, apartment, cottage, vineyard ownership, investment preferences, credit card usage, and internet shopping habits were selected as determinants of household consumption expenditures in the LAD-LASSO model. From the results of the study, it was found that machine learning algorithms can be used in the selection of the most appropriate variables in the course of the construction of microeconometric models.

Keywords
Machine Learning, LAD-LASSO Regression, Household Consumption Expenditures

JEL Classification
C31, C55, D12

Extended Summary

Household consumption expenditures play an important role both in providing information about the economic development levels of countries and determining rational production policies together with the determination of socioeconomic determinants. In literature, there were many studies on consumption expenditures. Although these studies aimed to select variables that determine consumption and obtain the most appropriate statistical and econometric model, these studies were modeled with different variables.

The Least Squares regression model (LS) is one of the most widely used estimation methods but LS estimators give unrealistic predictions in the presence of long-tailed errors, so LAD estimators are often used. However, since the number of variables in large data sets is high and the number of candidate models increases exponentially, the best model cannot be selected due to processing complexity. For this reason, Wang, Li, and Jiang (2007) developed the LAD-LASSO method, which enables the best model selection using the LASSO type penalty method, while obtaining robust estimators in the presence of outliers and long-tailed errors. The Household Budget Survey Consumption Expenditures dataset of Turkey contains both a great number of observations and many variables. Since the income distribution is not homogeneous in Turkey, household consumption expenditure does not show a homogeneous structure. Therefore, the LAD-LASSO, a penalty based machine learning method based on dimension reduction, was used in the analysis of the Household Budget Survey household data set in this study.

This study examined how the supervised machine learning methods help us to select the relevant variables of the Household Budget Survey Consumption Expenditures dataset with outliers in order to achieve a better performance in predicting and forecasting performances of the Household Consumption Expenditures Model. Since the main purpose of a penalty-based variable selection method is the only estimation

and causal and statistical inferences cannot be made from the obtained models, the results of the LAD-LASSO regression were evaluated in terms of variable selection and modeling.

In the study, the Household Consumption Expenditure model was predicted with the EKK method first, and diagnostic tests were applied to investigate the deviations from assumptions and outliers. To detect outliers, diagnostic tests were utilized to standard and student type residuals, and the presence of outliers was detected in 410 observations. In addition to the LASSO regression, the LAD and LAD-LASSO methods were predicted, which enabled robust estimators to be obtained in the presence of outliers and long-tailed errors; The results were compared and interpreted. The EKK and LASSO models prediction performance comparisons made use of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and R-Squared (R^2) criteria which gave very similar results.

According to the analyzed results, it was concluded that the LAD-LASSO machine learning method, which enables the selection of variables while obtaining robust predictors in the presence of long-tailed errors, is the most successful in prediction performance and forecasting accuracy. Several fundamental variables such as income, saving, and household size increased the household consumption expenditures for all models. In addition to these variables, other variables including the structure of a room, the kitchen, bathroom floors, heating, air conditioning preferences, energy sources used, detached house, apartment, cottage, vineyard ownership, investment preferences, credit card usage, and internet shopping habits were selected as determinants of household consumption expenditures in the LAD-LASSO model. From the results of the study, it was found that machine learning algorithms can be used in the selection of most appropriate variables in the course of the construction of microeconometric models. Although, penalty-based machine learning methods are successful methods in determining the model in data sets with a large number of variables, they should be used carefully because they make predictions based on correlation rather than causality.

Hanehalkı Tüketim Harcamalarının Mikroekonometrik Analizi: LAD-LASSO Yöntemi

Hanehalkı tüketim harcamaları makroekonomik bağlamda ülkelerin ekonomik gelişmişlik düzeyleri hakkında bilgi veren temel göstergelerden biridir (Varlamova ve Larionova,2015:727; Çalmaşur ve Kılıç,2018:61). Mikroekonomik bağlamda ise sosyoekonomik belirleyicilerinin tespit edilmesiyle beraber akılcı üretim politikalarının belirlenmesinde önemli bir rol oynamaktadır (TUİK,2018). Bu nedenle literatürde tüketim harcamalarını konu alan çok fazla sayıda çalışma bulunmaktadır. Bu çalışmalarda tüketimi belirleyen değişkenlerin tespit edilip en uygun istatistik ve ekonometrik modelin elde edilmesi amaçlansa da çalışmaların birbirinden farklı değişkenler ile modellenmiş olduğu görülmektedir.

Bir En Küçük Kareler (EKK) regresyonu modelinde ihmal edilmiş değişken hatası sapmalı tahmincilerin elde edilmesine, gereksiz bir değişkenin yer alması ise etkinlik kaybıyla beraber öngörü başarısının düşmesine neden olmaktadır. Ayrıca EKK tahmini için geliştirilmiş Akaike Bilgi Kriteri (AIC), Bayes Bilgi Kriteri (BIC) ve Schwartz Bilgi Kriteri (SIC) model seçim kriterlerinin hangisinin tercih edilmesi gerektiği ile ilgili genel bir kanı bulunmamaktadır (Shi ve Tsai, 2002:237). Uzun kuyruklu hataların varlığında EKK tahmincileri gerçeklikten uzak tahminler verdiğinden LAD tahmincileri sıklıkla kullanılmaktadır. Fakat büyük veri setlerinde değişken sayısı fazla olduğundan ve aday modellerin sayısı üssel arttığından işlem karmaşası nedeniyle en iyi modelin seçimi yapılamamaktadır (Wang, Li ve Jiang, 2007:1). Bu nedenle Wang, Li ve Jiang (2007) aşırı değerlere ve uzun kuyruklu hataların varlığında dirençli tahminciler elde edilirken LASSO tipi penaltı yöntemini kullanarak en iyi model seçimini yapan LAD-LASSO yöntemini geliştirmişlerdir. Türkiye'nin Hanehalkı Bütçe Anketi Hane veri seti hem çok sayıda değişken hem de çok sayıda gözlem sayısı içermektedir. Ayrıca gelir dağılımının homojen olmadığı Türkiye gibi ülkelerde hane halkı tüketim harcamaları da homojen bir yapı göstermemektedir. Bu nedenlerle bu çalışmada Hanehalkı Bütçe Anketi Hane veri setinin analizinde boyut indirgeme esaslı penaltı temelli makine öğrenme yöntemi olan LAD-LASSO kullanılmıştır. Literatürde hanehalkı araştırmasında makine öğrenme yöntemlerinin kullanıldığı çok fazla çalışma bulunmamaktadır. Mevcut çalışmalardan Gaffney ve Kirkby (2018), 1999-2003 yılları arasındaki PSID (Panel Study of Income Dynamics) veri setine klasik doğrusal regresyon yöntemin yanında Regresyon Ağacı, Topluluk öğrenmesi, Destek Vektör Makineleri ve LASSO regresyon olmak üzere çeşitli makine öğrenmesi yöntemleri uygulamış ve yaşam döngüsü sürekli gelir hipotezi (Life-Cycle Permanent-Income Hypothesis) denkleminin tüketimi tanımlamada en uygun fonksiyon olduğu sonucuna ulaşmışlardır. Andini vd. (2018) çalışmasında toplam yiyecek-içecek tüketimi harcamalarının tahmini için Italya Merkez Bankası (Bank of Italy)'nın 2014 yılı SHIW (hane halkı gelir ve refah düzeyi araştırması) veri setine regresyon ağacı, k-en yakın komşu ve rassal orman makine öğrenmesi yöntemlerinin

yanı sıra doğrusal olasılık modeli uygulamış ve regresyon ağacı yönteminin tahmin performansı açısından en uygun yöntem olduğu sonucuna ulaşmışlardır. Önder ve Turgut (2018) çalışmasında hanehalkı kiralık konut talebinin belirleyicilerini tespit etme amacıyla 2015 yılı Hanehalkı Bütçe Anket verilerine çeşitli makine öğrenmesi yöntemleri uygulanmış ve performans karşılaştırması sonucunda karar ağacı en uygun yöntem seçilmiştir. Obrizan vd. (2019) çalışmasında Macaristan'ın sigara tüketimi harcamalarını analiz etmek amacıyla GeoStat'tan elde edilen 2016 yılı IHS (Birleşik Hanehalkı Araştırması) veri setine EKK, rassal orman, gradyan arttırma (gradient boosting) ve derin öğrenme yöntemleri uygulamış ve örneklem dışı tahminleri karşılaştırmışlardır. Sonuç olarak makine öğrenmesi algoritmalarına kıyasla EKK'nın sınırlı bir tahmin performansı olduğu bulgusunu elde etmişlerdir. Azzopardi vd. (2019) çalışmasında ABD'nin hanehalkı finansal kırılganlığını ölçme amacıyla Federal Reserv Bankası (FED)'ın 1998, 2007 ve 2016 yılları SCF (Tüketici Finans Araştırması) veri setine HAC ve k-means kümeleme makine öğrenme yöntemleri uygulamış, 2016 yılındaki hanehalkı finansal kırılganlığının %28 ile 2007 ve 1998 yıllarından daha yüksek olduğu sonucunu elde etmişlerdir. Selim ve Demirkan (2020) hanehalkı gıda harcamalarını etkileyen faktörleri belirlemek amacıyla çalışmalarında yarı logaritmik regresyonun yanı sıra yapay sinir ağları yöntemini uygulamış, ön tahmin performansı karşılaştırması sonucunda yapay sinir ağları yönteminin tahmin performansının daha yüksek olduğu bulgusuna ulaşmışlardır.

Bu çalışmanın amacı Türkiye'nin Hanehalkı Tüketim Harcamalarını etkileyen faktörlerin belirlenebilmesi için EKK, LAD regresyonu klasik tahmin yöntemlerinin yanı sıra LASSO ve LAD-LASSO makine öğrenmesi yöntemlerini kullanılarak elde edilen modellerin tahmin ve öngörü performanslarını karşılaştırarak ilgili değişkenlerin yer aldığı en uygun modelin belirlenmesini sağlamaktır. Bilindiği gibi, penaltı temelli değişken seçme yöntemlerinin asıl amacı sadece tahmin olduğundan, elde edilen modellerden nedensel ve istatistiksel çıkarımlar yapılamamaktadır (Ahrens, 2019:2). Bu nedenle çalışmada özellikle LAD-LASSO regresyonu sonuçları değişken seçimi ve modelleme açısından değerlendirmeye alınmıştır.

Çalışmada hanehalkı tüketim harcamaları modeli ilk olarak EKK yöntemi ile tahmin edilmiş, varsayımlardan sapmalar ve aşırı değerlerin incelenmesi amacıyla diagnostik testler uygulanmıştır. Yöntem olarak LASSO regresyonun yanı sıra aşırı değerler ve uzun kuyruklu hataların varlığında robust (dayanıklı) tahmincilerin elde edilmesini sağlayan LAD ve LAD-LASSO yöntemleri ile tahmin edilerek tahmin performansları karşılaştırılmış; en uygun ve başarılı yöntem olan LAD-LASSO yönteminin sonuçları yorumlanmıştır.

Çalışmanın 2. bölümünde veri seti ve yöntem, 3. bölümünde Metodoloji ve 4. bölümünde ampirik bulgular ve 5. bölümünde sonuç kısmı yer almaktadır.

Veri Seti ve Yöntem

Bu çalışmada Türkiye İstatistik Kurumu (TUİK)'den elde edilen 2018 yılı Hanehalkı Bütçe Anketi Hane veri seti kullanılmıştır. Logaritmik hanehalkı tüketim harcaması değişkeni bağımlı değişken olarak alınmış ve belirleyicileri için veri seti ön işleme tabi tutulmuştur. Bu uygulamada hanehalkı tipi, konut tipi, zemin salon, zemin banyo, yakıt türü, tasarruf gibi bazı kategorik değişkenler kategori sayısı m olmak üzere (m-1) sayıda [0,1] kodlu kukla değişken olarak tanımlanmıştır. Konut piyasa değeri ve kişi başı gelir değişkenleri kullanılarak oluşturulan karşılıklı etkileşim değişkeni(KPD_KBG)[1],

$$KPD\_KBG = (Konut\_piyasa\_Deger)x(\frac{gelir}{HHB})$$

formülüyle hesaplanmış, (*Ev_Sahipliği*)x(*Konut_Borç*)[2] karşılıklı etkileşim ve kuadratik Hanehalkı büyüklüğü (HHB)[3] değişkenleri ile beraber veri setine eklenmiştir. Veri setinde yer alan gelir, kira tutarı, konut piyasa değeri, konut süre, konut alan, konut piyasa değer-kişi başı gelir (KPD_KBG) sürekli değişkenleri doğal logaritmaları alınarak açıklayıcı değişken olarak çalışmaya dahil edilmiştir. Eksik veriler veri setinden çıkarılmıştır. Bu işlemlerin sonucunda 132 değişken ve 11828 gözlemden oluşan veri setinden, 175 değişken ve 8490 gözlem ile büyük bir hanehalkı veri seti elde edilmiş sonrasında bu veri seti 80:20 oranında öğrenme ve test veri setlerine bölünmüştür. Öğrenme veri seti tahmin performansını, test veri seti ise öngörü başarısını ölçme amacıyla kullanılmıştır. Çalışmada öğrenme veri setine regresyon yöntemlerinden EKK ve LAD regresyonu, Makine öğrenmesi yöntemlerinden ise LASSO ve LAD-LASSO yöntemleri uygulanmıştır.

Metodoloji

En Küçük Mutlak Sapma Regresyonu (LAD), Roger Joseph Boscovich (1757) tarafından ortaya atılmış uzun kuyruklu hataların ve bağımlı değişkende aşırı değerlerin varlığında gerçeklikten uzak tahminler veren EKK regresyonu yerine kullanılan bir yöntemdir. Yöntemin fonksiyonu,

$$\min_{\beta} \sum_{i=1}^{n} |y_i - \sum_{j=o}^{p-1} x_{ij}\beta_j| \, , i = 1, \dots, n, j = 0, \dots, p-1$$

şeklindedir. Burada $y \epsilon R^n$ olmak üzere y_i, bağımlı değişken vektörü; $(x_{i1}, \dots, x_{ik})$ ise açıklayıcı değişken vektörüdür.

En küçük Mutlak Küçültme ve Seçim Operatörü (LASSO), ilk olarak Tibshirani (1996) tarafından ortaya atılmıştır. Belirli sabit bir katsayıdan daha küçük

1 Mian vd.(2013) çalışması temel alınarak hesaplanmış ve veri setine eklenmiştir.

2 Sec ve Zemcik (2007) çalışması temel alınarak veri setine eklenmiştir.

3 Showers ve Shotick(1994) çalışması referans alınarak veri setine eklenmiştir.

olan katsayıların mutlak değerlerinden yola çıkarak artıkların kareleri toplamını minimize eden bir değişken seçme yöntemidir. LASSO Regresyon yönteminin Lagrange formundaki fonksiyonu;

$$\hat{\beta}_{LASSO} = argmin(\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{i=1}^{m} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p}|\beta_j|)$$

olarak ifade edilebilir. Burada *n* gözlem sayısı, *m* parametre sayısı, *i* gözlem sayısı indisi, *j* parametre sayısı indisidir. Burada λ negatif olmayan ayarlama parametresidir.

LAD-LASSO yöntemi, Wang, Li ve Jiang (2007) tarafından bir ℓ_1 penaltı değişken seçme operatörü olan LASSO ile robust bir regresyon yöntemi olan LAD regresyonun birleştirilmesiyle geliştirilmiş bir yöntemdir. Kullanılan fonksiyon,

$$LAD - LASSO = L(\beta) = \sum_{i=1}^{n}|Y_i - X_i^{'}\beta| + n\sum_{j=1}^{p}\lambda_j|\beta_j|$$

şeklindedir. Burada *n* gözlem sayısı, *p* parametre sayısı, *i* gözlem sayısı indisi, *j* parametre sayısı indisidir. Burada λ negatif olmayan ayarlama parametresidir.

Ampirik Bulgular

Bağımlı değişkenin, bağımsız değişkenlerle doğrusal bir nedensellik ilişkisi olduğu varsayımı altındaki doğrusal regresyon modelleri yaygın olarak kullanılmaktadır. Klasik literatürde bağımlı değişken ile tahmini değerlerinden elde edilen hataların kareli toplamının minimizasyonuna dayanan ve bu nedenle literatürde En Küçük Kareler olarak adlandırılan EKK regresyonu en çok kullanılan regresyon yöntemidir. Tahmininde hata kareleri toplamının minimizasyonuna dayanan klasik optimizasyon yöntemleri kullanılmaktadır (Arthanari ve Dodge,1993:10). Bu hesaplanması kolay tahmin yöntemleri EKK regresyonunun daha çok tercih edilmesinin temel sebeplerinden biridir. EKK regresyonun başarılı sonuçlar verebilmesi için bazı temel varsayımların sağlanması gerekmektedir (Rao,1973:363). Klasik EKK yöntemi model tanımlama hatası, değişen varyans, çoklu doğrusal bağlantı, otokorelasyon gibi bazı temel varsayımlar altında ilgili değişkenlerin seçiminde nedensellik yaklaşımı açısından oldukça başarılı bir yöntemdir. Bu nedenle bu çalışmada öğrenme veri setine öncelikle doğrusal regresyon yöntemi EKK uygulanmış, elde edilen model sonuçlarına değişen varyans ve normal dağılım testleri yapılmış, modelin regresyon temel varsayımlarını sağlamadığı tespit edilmiştir. EKK regresyonunda değişken seçimi neden sonuç ilişkisine göre yapılsa da değişken seçiminde kullanılan altküme seçicisi (subset selection) yöntemi bir değişken ekleme çıkartma prosedürü olduğundan bazı yanıltıcı sonuçlara neden olabilmektedir. Son zamanlarda hem regresyon temel varsayımlarının sağlanması kısıtlaması hem de değişken ekleme

çıkartma prosedürünün parametre seçiminde güvenilir sonuçlar vermemesi nedeniyle istatistikçi ve ekonometrisyenler değişken seçme ve düzenleme ile yüksek tahmin ve öngörü başarısı elde edilmesine olanak sağlayan penaltı temelli regresyon yöntemlerini sıklıkla kullanmaktadırlar. Bu sebeple bu çalışmada öğrenme veri setine Tibshirani (1996) tarafından önerilen LASSO regresyon yöntemi uygulanmıştır. Her iki modelin öğrenme veri seti ile tahminci başarısı, test veri seti ile tahmincilerin örneklem dışı öngörü başarısı RMSE (Hataların Kareli Ortalamasının Karekökü), MAE (Ortalama Mutlak Hata) ve R^2 karşılaştırma kriterleri ile ölçülmüştür. Performans karşılaştırmalarında RMSE, MAE kriterlerinin minimum ve R^2'nin maksimum değer aldığı model en başarılı model olarak belirlenmiştir. Model karşılaştırma sonuçları Tablo 1'de yer almaktadır.

Tablo 1.

Tahmin ve Öngörü Başarısı Tablosu

	Öğrenme Veri Seti			*Test Veri Seti*		
Yöntem	*RMSE*	*MAE*	R^2	*RMSE*	*MAE*	R^2
EKK	0,4702	0,3126	0,23	0,4793	0,3157	0,21
LASSO	0,4655	0,3085	0,25	0,4778	0,3140	0,21

Tablo 1 incelendiğinde, hatalara dayalı karşılaştırma kriterlerine göre LASSO ve EKK tahmin performanslarının oldukça birbirine yakın olduğu görülmektedir. Uygun EKK modelinin tahmin sonuçları Tablo 2'de yer almaktadır.

Tablo 2

EKK Regresyonu Tahmin Sonuçları

Değişken	*Katsayı*	*Std. Hata*	*Robust std. Hata*	*t- test istatistiği*	*P>|t|*
HH_TİP_HBA1	-0,1146	0,2179	0,0246	-4,6600	0,0000
ZEMİN_BANYO_6	-0,0943	0,0204	0,0232	-4,0600	0,0000
KÖMÜR	-0,0727	0,0151	0,0149	-4,8900	0,0000
TEZEK1	-0,3315	0,0484	0,0753	-4,4000	0,0000
SAUNA	0,3306	0,1578	0,0363	9,1000	0,0040
ASANSOR	0,0379	0,0167	0,0156	2,4200	0,0150
GARAJ	0,0580	0,0214	0,0187	3,1000	0,0020
COCUK_OYUN_ALAN	-0,0843	0,0194	0,0222	-3,7900	0,0000
CEP_TELEFON_SAYI	0,0312	0,0067	0,0071	4,3900	0.0000
PANEL_TV_SAYI	0,0389	0,0103	0,0101	3,8700	0,0000
BULAŞIK_MAKİNE_SAYI	0,0577	0,0149	0,0150	3,8500	0,0000
CAMAŞIR_MAKİNE_SAYI	0,1402	0,0434	0,0589	2,3800	0,0170
KLİMA_SAYI	0,0319	0,0111	0,0111	2,8700	0,0040
OTOMOBİL_SAYI	0,1175	0,0112	0,0108	10,8800	0,0000
MOTOSİKLET_SAYI	0,0788	0,0202	0,0173	4,5600	0,0000
TARLA_SAHİPLİĞİ	0,0492	0,0160	0,0156	3,1500	0,0020
TARLA_KİRAYA_VERİLEN	-0,0023	0,0011	0,0012	-1,9400	0,0530
ALKOL_ALIŞKANLIK	0,0750	0,0256	0,0220	3,4100	0,0010
DIŞARIDA_YEMEK	0,0550	0,0129	0,0123	4,4500	0,0000
DERGİ_ALIŞKANLIK	0,0823	0,0383	0,0299	2,7500	0,0060
SİNEMA_ALIŞKANLIK	0,0657	0,0217	0,0195	3,3700	0,0010

ÜCRETLİ_TV_ALIŞKANLIK	0,0644	0,0190	0,0174	3,7100	0,0000
ÜCRETLİ_SPOR_ALIŞKANLIK	0,0600	0,0219	0,0182	3,3000	0,0010
KAHVEHANE_ALIŞKANLIK	0,0704	0,0137	0,0127	5,5200	0,0000
KREDİ_KARTI	0,0615	0,0134	0,0130	4,7500	0,0000
İNTERNET_ALIŞVERİŞ_SIKLIK	0,0312	0,0115	0,0103	3,0400	0,0020
SQ_HHB	0,0010	0,0004	0,0005	1,9200	0,0550
GELİR	0,0897	0,0130	0,0135	6,6500	0,0000
SABİT	4,3571	0,0930	0,1047	41,6200	0,0000
VIF skoru				1,2500	
Breusch-Pagan / Cook-Weisberg Değişen Varyans Testi				χ^2 (1)=366.5200	0,0000
Shapiro-Wilk Normallik Testi				Z=17.3790	0,0000
Kolmogorov-Smirnov Normallik Testi				D = 0.2105	0,0000
***** Breusch-Pagan / Cook-Weisberg Değişen Varyans Testi temel hipotezi *"Ho: Sabit varyans varsayımı geçerlidir"*, Shapiro-Wilk ve Kolmogorov- Smirnov normallik testlerinin temel hipotezi ***"****Ho: Hata terimleri dağılımı normal dağılıma uygunluk göstermektedir"* olarak kurulmaktadır.					

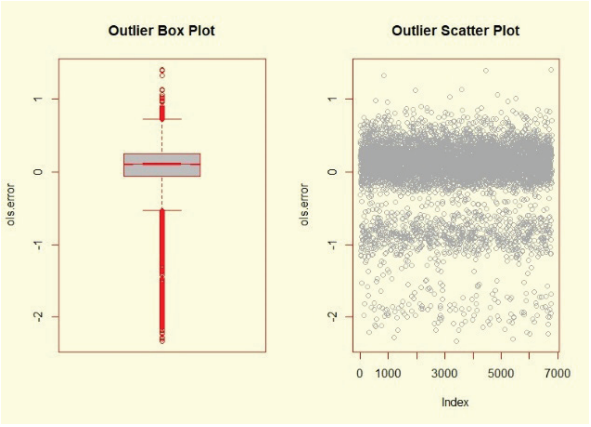
Tablo 2 incelendiğinde analize alınan 174 değişken arasından 28 değişkenin katsayıları istatistiksel olarak anlamlı bulunmuştur ve modele dahil edilmiştir. Gelir temel değişkeninin yanı sıra hane halkı büyüklüğü yerine kuadratik hanehalkı büyüklüğü değişkeninin modelde yer aldığı görülmektedir. Elde edilen sonuçlara bakıldığında Sauna sahipliği (SAUNA) değişkeni harcama üzerinde en büyük pozitif etkiye sahip değişken olarak bulunurken[0.3306], tezek kullanımı (TEZEK1) değişkeninin hane halkı harcamaları üzerinde en negatif etkiye sahip değişken olduğu[-0.3315] görülmektedir. Hanenin kömür kullanımı (KÖMÜR) ve banyonun zemininin kara beton olması (ZEMİN_BANYO_6) değişkenlerinin de harcamalar üzerinde azaltıcı bir etkiye sahip olduğu görülmektedir. Bunların dışında tarla sahipliği (TARLA_SAHİPLİĞİ) değişkeni harcamalar üzerinde pozitif bir etkiye sahipken, hanehalkının kiraya verilen tarlasının olmasının (TARLA_KİRAYA_VERİLEN) harcamalar üzerinde tam tersi bir etkiye sahip olduğu görülmektedir. Tablodaki bilgiler doğrultusunda bu değişkenlerin dışında kalan tüm değişkenlerin harcamalar üzerinde arttırıcı bir etkiye sahip olduğu bulunmuştur.

Model için sabit varyans varsayımın geçerliliğini incelemek için Breusch-Pagan / Cook-Weisberg Değişen Varyans Testi uygulanmıştır. Breusch-Pagan/Cook-Weisberg [chi2(1)=366,52; p=0,00] testi sonucunda değişen varyans varsayımına karşılık sabit varyans olduğu varsayımını savunan temel hipotez reddedilmiş ve değişen varyans olduğu bulgusuna ulaşılmıştır. Bu nedenle katsayılar robust standart hatalar kullanılarak yeniden tahmin edilmiştir.

Modelde çoklu doğrusal bağlılık analizi için ise VIF değerleri hesaplanmıştır. Hesaplanan VIF değerleri 5 değeri ile karşılaştırıldığında, tüm VIF değerlerinin 5'ten küçük olduğu görülmüş ve modelde yüksek derecede çoklu doğrusal bağlılık olmadığı sonucuna ulaşılmıştır[4].

4 Değişken bazındaki VIF skor tablosu yazarlardan temin edilebilir.

Normallik varsayımı için Shapiro-Wilk ve Kolmogorov-Smirnov testleri uygulanmıştır. Shapiro-Wilk(1965) normallik testi [Z=17,38; p=0,00] test istatistiği ile ve Kolmogorov-Smirnov normallik testi [D = 0,2105; p=0,00] sonucunda hataların normal dağılmadığı görülmektedir. Bir EKK tahmininde hataların normal dağıldığı varsayımının sağlandığı koşulda tahminciler arasından minimum varyanslı tahminci seçilir. Hatalar normal dağılmıyorsa EKK tahmincisinin varyansı minimum olası tahminci varyansından daha büyük olabilmektedir. Böylece EKK tahmincileri ve testleri etkinlik kaybına uğrayabilir ve bazı aşırı değerlerin varlığı sapmalı ve etkin olmayan EKK sonuçlarının elde edilmesine neden olabilir (Birkes ve Dodge, 1993:190). Burada aşırı değerlerin varlığı regresyon temel varsayımlarının sağlanamamasına neden olabileceğinden, standart ve student türü artıklar incelenmiş ve diagnostik testlerin sonunda 410 gözlemde aşırı değerin varlığı kanıtlanmıştır[5]. Aşırı değer kutu ve dağılım grafikleri Şekil 1'de yer almaktadır.



Şekil 1. Aşırı Değer Kutu ve Dağılım Grafikleri

Şekil 1'de aşırı değer kutu ve dağılım grafikleri yer almaktadır. Kutu grafiği incelendiğinde hataların normal dağılmadığı ve sağa çarpık bir dağılım sergilediği söylenebilir.

Aşırı değer dağılım grafiği incelendiğinde hataların [0,1] aralığında daha yoğun olduğu ve aşırı değerlerin [0, -2] aralığına doğru kuyruklu bir yayılım sergilediği açıkça görülmektedir.

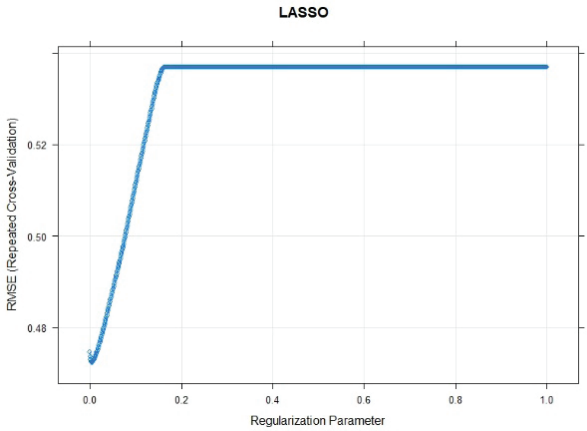
Parametrik yöntemle en iyi modele ulaşmada sorun olan aşırı değerlerin etkisini en aza indirgeyecek şekilde veriye en uygun olan yapının belirlenmesi gerekmektedir. Bu durumda birçok veri noktasında etkili olan ve toplam hata kare değerinde artmaya neden olan aşırı değerleri belirleyerek etkinlik kaybına neden olabilecek kaldıraç noktalarını tespit edebilen robust tahmin yöntemleri kullanılır (Hampel vd., 1986:11). Bu tahmin yöntemlerinden birisi ise mutlak hataların ortalamasının minimizasyonuna

5 Aşırı değer test sonuçları yazardan temin edilebilir.

dayanan ve literatürde en küçük mutlak sapma olarak adlandırılan LAD regresyonudur. LAD regresyonu ℓ_1-norm minimizasyonu olarak da bilinmektedir. Bu regresyonun çözümünde konum parametresi olarak ortalama yerine medyan kullanılmaktadır. LAD regresyonu, özellikle EKK regresyonundan elde edilen tahmin hataların Cauchy veya Laplace gibi uzun kuyruklu dağılıma sahip olduğu durumlarda, veri setlerinde bulunan aşırı değerlere karşı dayanıklı tahmincilerin elde edildiği bir yöntem olarak kullanılmaktadır. Bunun yanında yöntem bağımlı değişkende var olan dikey aşırı değerleri yakalamada da oldukça başarılıdır (Dodge, 1997: 145). Robust regresyonda değişken seçimi EKK regresyonuna benzer bir şekilde bazı parametrelerin sıfıra eşit olup olmadığının test edilmesiyle gerçekleştirilir. Fakat ki-kare χ^2 istatistiklerine dayanan uyum iyiliği ölçüleri, χ^2 istatistikleri büyük veri setlerinde anlamlılık seviyesini aşma eğilimi gösterdiğinden araştırmacıları yüksek bir değerle temel hipotezi kabule yöneltebilmektedir (Ylvisaker, 1977:553). Bu nedenle robust uyum iyiliği ölçüleri model belirlenirken büyük veri setlerinin değişken seçiminde güvenilir sonuçlar vermemektedir.

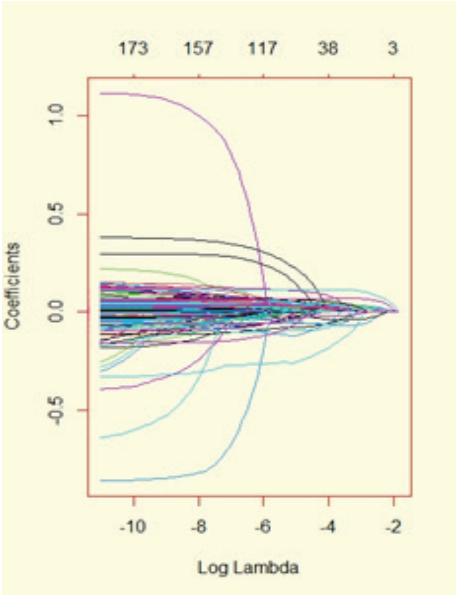
Bağımlı değişkeni etkileyen özelliklerin sayısı arttıkça LAD ve EKK gibi doğrusal regresyon yöntemleri yeterli olmayabilir. Doğrusal regresyonda her özellik için ayrı bir parametre oluşturulmaktadır. Modelde çok fazla parametre olduğunda olası modellerin sayısı üssel arttığından oluşabilecek işlem karmaşası nedeniyle değişken seçiminde ve katsayı tahminlerinde katsayı büyüklüklerinin penaltı edilmesine dayanan LASSO regresyonu yaygın olarak kullanılmaktadır. Fakat LASSO regresyonu minimum varyanslı parametre seçimini ortalamaya dayalı tahmincilerle elde edilen hataların kareli ortalamasının minimizasyonuna dayalı olarak yaptığından aşırı değerlerin ve uzun kuyruklu hataların varlığında sapmalı ve etkin olmayan tahmincilerin elde edilmesine neden olabilmektedir. Bu nedenle, bu çalışmada veri setine, aşırı değerlerle kirlenmiş büyük veri setlerinde ve hataların normal dağılmama durumunda literatürde yaygın olarak kullanılan ortalama yerine medyana dayalı tahminciler ile başarılı sonuçlar elde edilmesini sağlayan LAD regresyonunun yanı sıra ℓ_1- penaltı değişken seçme yöntemi LASSO ve LAD regresyonun birleşimi olan LAD-LASSO yöntemleri uygulanarak parametre tahmincileri elde edilmiştir.

LASSO ve LAD-LASSO penaltı temelli makine öğrenme yöntemlerinde aşırı uyum sorununun önüne geçebilmek için sıfıra eşit olmayan katsayıların penaltı edilmesi ile elde edilen lambda (λ) hiper-parametresi kullanılmaktadır. Hiper-parametreler varyans-sapma dengesini ayarlayarak modelin iyi fitlenmesini sağlar. Bu nedenle seçimi çok önemlidir (Pedregosa, 2016:739). Şekil 2 LASSO için çeşitli λ hiper-parametrelerin değişimlerine karşılık 10-fold CV (10 katmanlı çapraz geçerlilik)-RMSE değişimi gösterilmektedir.



Şekil 2. LASSO Ayarlama Parametresi

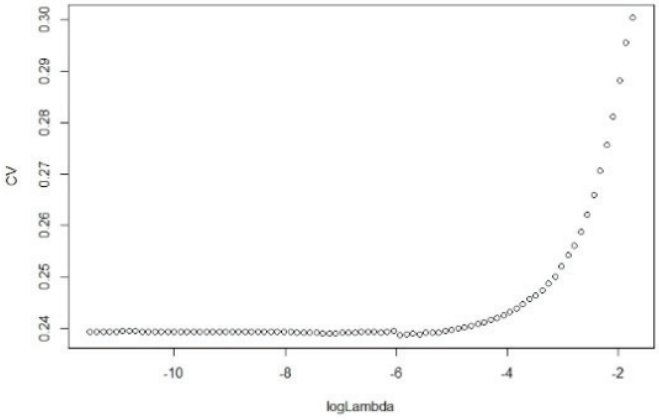
Şekil 2'de yer alan 10-fold CV-RMSE değerleri λ hiper-parametresinin bir fonksiyonudur. Çalışmamızda kullandığımız LASSO modelinde penaltı düzeyi arttıkça CV-RMSE değerlerinin de arttığı açıkça görülmektedir. Burada penaltı temelli yöntemlerin hiper-parametre optimizasyonunda λ seçimi 10-fold CV yönteminin kullanılması sonucu CV-hatalarını minimize eden λ değerlerine bakılarak gerçekleştirilmiştir ve optimum λ değeri 10-fold CV-RMSE değerinin minimum olduğu noktada [0,00401] olarak elde edilmiştir.



Şekil 3. LASSO Katsayı Grafikleri

Şekil 3'te katsayıların tahminci skorları log(λ)'nın bir fonksiyonu olarak gösterilmektedir. Grafikte log(λ)'nın artan değerlerine karşılık katsayılardaki küçülme

ve parametre uzayından sıfıra eşit olmayan katsayılı değişkenlerle oluşturulmuş aralıklı model yapısına geçiş açıkça görülmektedir. Bu çalışmada kullanılan LASSO modelinin 174 açıklayıcı değişken arasından seçilen katsayısı sıfıra eşit olmayan değişkenlerinin sayısı 100 olarak elde edilmiştir.



Şekil 4. LAD-LASSO Ayarlama Parametresi

Şekil 4'de LAD-LASSO log(λ) değerlerinin değişimlerine karşılık CV hatalarının değişimi gösterilmektedir. Optimum λ değeri CV hatalarının minimum olduğu noktada [0,00266] olarak elde edilmiştir.

Bu çalışmada Türkiye'nin Hanehalkı Tüketim Harcamaları modeli öncelikle EKK regresyonu ile tahmin edilmiş fakat yapılan diagnostik testlerin sonucunda aşırı değerlerin olduğu tespit edilmiştir. Tahmincilerin sapmalı olmaması ve etkinlik kaybına uğramaması için veriye en uygun olan yapının belirlenmesi amacıyla aşırı değerlerin etkisini en aza indirgeyerek birçok veri noktasında etkili olan ve etkinlik kaybına neden olabilecek kaldıraç noktalarını tespit edebilen LAD ve LAD-LASSO yöntemleri uygulanmış ve sonuçlar Tablo 3'de sunulmuştur.

Tablo 3

Tahmin ve Öngörü Başarısı Tablosu

	Öğrenme Veri Seti			*Test Veri Seti*		
Yöntem	*RMSE*	*MAE*	R^2	*RMSE*	*MAE*	R^2
LAD	0,5016	0,3018	0,17	0,5043	0,2982	0,16
LAD-LASSO	0,4858	0,2816	0,23	0,4909	0,2832	0,21

Her iki yöntem için örneklem içi tahminler öğrenme veri seti, örneklem dışı tahminler test veri seti kullanılarak hesaplanmış; hatalara dayalı RMSE, MAE ve R^2 kriterleri hesaplanarak performans karşılaştırmaları yapılmıştır. Elde edilen bulgular ışığında LAD-LASSO'nun tahmin performansı [0,4858; 0,2816;0,23] skorlarıyla LAD regresyonundan yüksek bulunmuştur. Yöntemlerin öngörü başarıları karşılaştırıldığında [0,4909;0,2832;0,21] değerleriyle LAD-LASSO yönteminin

LAD regresyonundan daha başarılı olduğu bulgusuna ulaşılmıştır. Hem tahmin performansı hem de öngörü başarısı açısından en başarılı yöntem olarak belirlenen LAD-LASSO tahmin sonuçları Tablo 4'de sunulmuştur.

Tablo 4

LAD Regresyonu Tahmin Sonuçları

Değişken	*Katsayı*	*Std, Hata*	*t- test istatistiği*	*P>|t|*
SABİT	4,4883	0,0746	60,1763	0,0000
HH_TİP_HBA1	-0,1616	0,0261	-6,1877	0,0000
HH_TİP_HBA2	0,0450	0,0249	1,8073	0,0708
HH_TİP_HBA3	0,0762	0,0258	2,9508	0,0032
KONUT_TİP1	0,0131	0,0040	3,2668	0,0011
KİRA_SEKLİ	0,0176	0,0076	2,3154	0,0206
ODA_SAYI	0,0126	0,0050	2,5298	0,0114
KONUT_ALAN	0,0836	0,0357	2,3382	0,0194
KOMBİ	0,1005	0,0090	11,1153	0,0000
MERKEZİ_KALORİFER_BİNA	0,1188	0,0139	8,5172	0,0000
MERKEZİ_KALORİFER_BİNADIŞ	0,1664	0,0343	4,8448	0,0000
ELEKTRİK1	0,0798	0,0160	4,9750	0,0000
SICAKSU_ODUN	-0,0470	0,0258	-1,8207	0,0687
ASANSOR	0,0301	0,0096	3,1198	0,0018
COCUK_OYUN_ALAN	-0,0322	0,0104	-3,1103	0,0019
PANEL_TV_SAYI	0,0575	0,0053	10,9611	0,0000
BUZDOLABI_SAYI	0,1322	0,0086	15,4551	0,0000
MIKRODALGA_FIRIN_SAYI	0,0269	0,0079	3,4165	0,0006
CAMASIR_MAKİNE_SAYI	0,0943	0,0215	4,3758	0,0000
CAMASIR_KURUTMA_SAYI	0,0457	0,0223	2,0483	0,0406
İŞVEREN_ARAC_SAYI	0,0530	0,0278	1,9100	0,0562
DENİZTASIT_SAYI	0,1587	0,0260	6,1083	0,0000
APARTMAN_KİRAYA_VERİLEN	0,0456	0,0109	4,1868	0,0000
YAZLIK_SAHİPLİGİ	0,0768	0,0190	4,0410	0,0001
DUKKAN_SAHİPLİGİ	0,0562	0,0149	3,7691	0,0002
DERGİ_ALIŞKANLIK	0,0915	0,0323	2,8293	0,0047
SİNEMA_ALIŞKANLIK	0,1006	0,0127	7,9510	0,0000
ÜCRETLI_TV_ALIŞKANLIK	0,0810	0,0105	7,6856	0,0000
KAHVEHANE_ALIŞKANLIK	0,0782	0,0069	11,3928	0,0000
PAZAR_ALIŞKANLIK	0,0325	0,0068	4,7722	0,0000
T1_ALTIN	-0,0445	0,0190	-2,3412	0,0193
GELİR	0,0430	0,0046	9,3137	0,0000

Tablo 4 LAD regresyonu tahmin sonuçlarını göstermektedir. Veri setinde yer alan 174 açıklayıcı değişkenin katsayılarının 31 tanesi istatistiksel olarak anlamlı bulunmuş ve modele alınmıştır.

Tablo 5

LAD-LASSO Regresyonu Tahmin Sonuçları

Değişken	*Katsayı*	*Değişken*	*Katsayı*
SABİT	4,7941	DENİZTAŞIT_SAYI	0,0757
KONUT_TİP1	-0,0025	BALKON	0,0188
HH_TİP_HBA1	-0,0778	TELEFON_HAT_SAYI	0,0078
HH_TİP_HBA3	0,0072	CEP_TELEFON_SAYI	0,0251
KİRACİ	-0,0038	BİLGİSAYAR_SAYI	0,0062
KONUT_PİYASA_DEGER	0,0033	PANEL_TV_SAYI	0,0299
KONUT_SÜRE	-0,0062	VIDEO_KAMERA_SAYI	0,0089
ZEMİN_SALON1	0,0150	UYDU_ANTEN_SAYI	-0,0161
ZEMİN_SALON2	0,0043	CAMAŞIR_KURUTMA_SAYI	0,0009
ZEMİN_SALON4	0,0003	HALI_YIKAMA_SAYI	-0,0009
ZEMİN_SALON7	0,0170	OTOMOBİL_SAYI	0,0893
KONUT_ALAN	0,0264	BULAŞIK_MAKİNE_SAYI	0,0204
ODA_SAYI	0,0062	MİKRODALGA_FIRIN_SAYI	0,0006
ZEMİN_ODA1	0,0045	KLİMA_SAYI	0,0417
ZEMİN_ODA4	0,0159	İŞVEREN_ARAC_SAYI	0,0317
ZEMİN_ODA8	0,0775	MOTOSİKLET_SAYI	0,0466
ZEMİN_MUTFAK1	0,0038	MÜSKONUT_SAHİPLİĞİ	-0,0168
ZEMİN_MUTFAK2	-0,0189	MÜSKONUT_MİKTAR	0,0111
ZEMİN_MUTFAK7	-0,0227	MÜSKONUT_KİRAYA_VERİLEN	0,0051
ZEMİN_BANYO3	0,0240	DERİN_DONDURUCU_SAYI	0,0009
ZEMİN_BANYO6	-0,0133	APARTMAN_SAHİPLİĞİ	0,0043
ZEMİN_BANYO8	0,0437	APARTMAN_KİRAYA_VERİLEN	0,0214
KÖMÜR	-0,0009	APARTMAN_AYLIK_KİRA	0,00001
KONUTİÇİ_BAGIMSIZ_KLİMA	-0,0018	YAZLIK_SAHİPLİĞİ	0,0036
MERKEZİ_KALORİFER_BİNA	0,0179	TARLA_SAHİPLİĞİ	0,0195
MERKEZİ_KALORİFER_BİNADIŞ	0,0042	TARLA_KİRAYA_VERİLEN	-0,0017
KONUTİÇİ_MERKEZİ_KLİMA	-0,0369	BAĞ_MİKTAR	0,0009
ODUN2	0,0022	BAĞ_KİRAYA_VERİLEN	0,0028
KÖMÜR1	-0,0072	ARSA_MİKTAR	0,00001
KÖMÜR2	-0,0052	KREDİ_KARTI	0,0565
MUTFAK_ELEKTRİK	-0,0103	ÖZEL_SİGORTA	0,0487
MUTFAK_KÖMÜR	0,0025	ALKOL_ALIŞKANLIK	0,0470
TEZEK1	-0,1388	DÜKKAN_SAHİPLİĞİ	0,0292
TEZEK2	-0,0794	DÜKKAN KİRAYA_VERİLEN	-0,0086
SICAKSU_LPG	-0,0431	SİNEMA_ALIŞKANLIK	0,0370
SICAKSU_ELEKT	0,0120	KAHVEHANE_ALIŞKANLIK	0,0453
MUTFAK_ODUN	-0,0081	SİGARA_ALIŞKANLIK	0,0318
SICAKSU_KÖMÜR	0,0068	DERGİ_ALIŞKANLIK	0,0271
SICAKSU_TEZEK	-0,0330	PAZAR_ALIŞKANLIK	0,0031
ÖLÇEK_POSTA	-0,0046	ÜCRETLİ_TV_ALIŞKANLIK	0,0451
KONUT_İKİNCİ	0,0155	ELEKTRİKLİ_BİSİKLET_SAYI	-0,0085
TABAN_ISITMA	0,0223	GAZETE_ALIŞKANLIK	0,0169
SAUNA	0,1050	ALETLİ_SPOR_ALIŞKANLIK	0,0238
KABLO_YAYIN	0,0010	DIŞARIDA_YEMEK	0,0446
TUVALET	0,0264	İNTERNET_ALIŞVERIŞ_SIKLIK	0,0275
ÇOP_ÖĞÜTÜCÜ	0,0144	T1_GAYRIMENKUL	-0,013
KALORİFER	0,0334	T1_DÖVİZ	0,0563
COCUK_OYUN_ALAN	-0,0136	T1_KOOPERATİF	0,0091

ASANSOR	0,0234	T1_ALTIN	-0,0036
İNTERNET_SAYI	0,0085	T1_BANKA	0,0045
SICAKSU	0,0211	T1_FON_KATİLİM	0,0137
BUZDOLABI_SAYI	0,0411	T1_YAPMIYOR	0,0225
DOĞALGAZ	0,0216	T2_FON_KATILIM	0,0307
CAMAŞIR_MAKİNE_SAYI	0,0478	T2_İŞ_YATIRIM	0,0343
GARAJ	0,0124	SQ_HHB	0,0011
HAVUZ	0,0183	KPD_KBG	0,0005
GÜVENLIK_SİSTEM	0,0046	GELİR	0,0188

Tablo 5, LAD-LASSO regresyon sonuçlarını göstermektedir. Türkiye'nin Hanehalkı Tüketim Harcamaları modeli, tahmin ve değişken seçimi sonucunda 174 açıklayıcı değişken arasından katsayısı sıfıra eşit olmayan 113 ilgili değişken seçilerek oluşturulmuştur. Sonuçlar incelendiğinde LAD-LASSO modelinde yer alan 25 değişkenin aynı zamanda LAD regresyon modelinde de yer aldığı görülmektedir.

Sonuç

Hanehalkı tüketim harcamaları mikroekonomik bağlamda endüstriyel politikaların oluşturulmasında önemli bir kaynak teşkil ettiğinden tüketimi belirleyen en uygun değişkenlerin yer aldığı ekonometrik modeli belirlemek önem arz etmektedir. Bu çalışmada, hanehalkı tüketim harcamalarının ilgili değişkenlerinin belirlenmesi amacıyla TUİK'in 2018 Hanehalkı bütçe anketi hane veri setine EKK, LAD regresyonu klasik tahmin yöntemlerin yanı sıra LASSO ve LAD-LASSO makine öğrenmesi yöntemleri uygulanarak 4 farklı model oluşturulmuş ve en uygun modelin tespit edilmesi amacıyla yöntemler arasında performans karşılaştırmaları yapılmıştır.

İlk olarak hanehalkı veri setine EKK regresyonu yöntemi uygulanmıştır. EKK regresyonunda kullanılan altküme seçim (subset selection) değişken seçme yönteminin bir değişken ekleme-çıkarma prosedürü olması ve çok sağlıklı sonuçlar vermemesi nedeniyle veri setine akabinde penaltı temelli değişken seçme yöntemlerinden LASSO regresyon yöntemi uygulanmıştır. EKK ve LASSO modelleri için RMSE, MAE ve R^2 kriterlerine bakılarak yapılan tahmin ve öngörü performans karşılaştırmalarının birbirine çok yakın sonuçlar verdiği tespit edilmiştir. Ayrıca uygulanan testler sonucunda regresyon temel varsayımlarından normallik ve değişen varyans varsayımlarının sağlanmadığı bulgusuna ulaşılmıştır. Aşırı değerlerin yer aldığı hanehalkı veri setine uygulanan EKK regresyonu uzun kuyruklu hataların varlığında gerçeklikten uzak sonuçlar verebileceğinden aşırı değerlerden şüphelenilmiştir. Aşırı değerlerin saptanması amacıyla standart ve student tipi artıklara diagnostik testler uygulanmış ve 410 gözlemde aşırı değerlerin varlığı tespit edilmiştir. Bu tespit sonucuda veri setine aşırı değerlerin etkisini en aza indirgeyerek daha sağlam (robust) tahminciler elde edilmesini sağlayan LAD ve LAD-LASSO yöntemleri uygulanarak elde edilen modellerin hem tahmin hem de öngörü performansları karşılaştırılmıştır. RMSE, MAE ve R^2 kriterlerine göre yapılan karşılaştırmalar sonucunda LAD-LASSO

regresyonunun Türkiye'nin Hanehalkı Tüketim Harcamaları'nın ilgili değişkenlerini belirlemede en uygun yöntem olduğu sonucuna ulaşılmıştır.

Tahmin sonuçları değerlendirildiğinde harcamanın en temel belirleyicisi olan gelirin yanı sıra hanehalkı_tip_1, çocuk_oyun_alan, asansör, çamaşır_makine_sayi, Panel_TV_sayı, sinema_alışkanlık, kahvehane_alışkanlık, ücretli_TV_alışkanlık, dergi_alışkanlık değişkenleri 4 modelde ortak değişkenler olarak yer almaktadır. Bu değişkenlere ek olarak hanenin oda, mutfak, banyo zeminlerinin yapısı, ısınma, iklimlendirme tercihleri, kullanılan enerji kaynakları, müstakil konut, apartman, yazlık, bağ sahipliği ve yatırım tercihlerini içeren çeşitli değişkenler, kredi kartı kullanımı, internet alışveriş alışkanlığı değişkenlerinin yanı sıra verisetine sonradan eklenen kuadratik hanehalkı büyüklüğü ve konut piyasa değer-kişibaşı gelir karşılıklı etkileşim değişkenlerinin Türkiye'nin Hanehalkı Tüketim Harcamaları'nın belirleyicisi olarak LAD-LASSO modeline seçilmiş diğer değişkenler olduğu sonucuna ulaşılmıştır.

Elde edilen bulgular ışığında LAD-LASSO modelinin 25 ortak değişkenle 31 değişkenden oluşan LAD regresyon modelini büyük oranda kapsadığı söylenebilir. Bu sonuç da bizi ilgili değişkenlerin seçiminde hem tahmin hem de bir değişken seçme prosedürü olan LAD-LASSO'nun optimizasyon sonucu elde edilen ayarlama parametresi ile gerçekleştirdiği penaltı sayesinde, EKK ve LAD regresyon yöntemlerinden daha iyi sonuç verdiğini göstermiştir. Fakat penaltı temelli makine öğrenmesi yöntemleri değişken sayısının fazla olduğu veri setlerinde modeli belirlemede başarılı yöntemler olsalar da nedensellik yerine korelasyona dayalı tahminler yapmaları sebebiyle dikkatli kullanılmaları gerekmektedir.

Hakem Değerlendirmesi: Dış bağımsız.
Çıkar Çatışması: Yazarlar çıkar çatışması bildirmemiştir.
Finansal Destek: Yazarlar bu çalışma için finansal destek almadığını beyan etmiştir.

Peer-review: Externally peer-reviewed.
Conflict of Interest: The authors have no conflict of interest to declare.
Grant Support: The authors declared that this study has received no financial support.

Kaynakça/References

Ahrens A., Hansen, C. B., & Schaffer, M.E.(2019). Lassopack: Model Selection and Prediction with Regularized Regression in Stata. *IZA Institute of Labor Economics,* IZA DP No.12081.

Andini, M., Ciani, E., De Blasio, G., D'ignazio, A., & Salvestrini, V. (2018). Targeting with Machine Learning: An Application to A Tax Rebate Program in Italy. *Journal of Economic Behavior and Organization*, 156, 86–102.

Arthanari, T. S., & Dodge, Y. (1993). Mathematical Programming in Statistics. John Wiley&Sons, Inc., New York.

Azzopardi, D., Fareed, F., Lenain, P., & Shutherland, D. (2019). Assessing Household Financial Vulnerability: Empirical Evidence from the U.S. using Machine Learning. OECD Economic Survey of the United States: Key Research Findings 2019, 121-142.

Birkes, D. & Dodge,Y. (1993). Alternative Methods of Regression. John Wiley&Sons, Inc., New York.

Breusch, T. S., & Pagan, A. R. (1979). A Simple Test for Heteroskedasticity and Random Coefficient Variation. *Econometrica, 47*(5), 1287-1294.

Cook, R. D. & Weisberg, S. (1983). Diagnostics for Heteroskedasticity in Regression. *Biometrika, 70*(1), 1-10.

Çalmaşur, G. &Kılıç, A. (2018). Türkiye'de Hanehalkı Tüketim Harcamalarının Analizi. *ETÜ Sosyal Bilimler Enstitüsü Dergisi,* 5, 61-73.

Dodge, Y. (1997). LAD Regression for Detecting Outliers inResponse and Explanatory Variables. *Journal of Multivariate Analysis*, 61, 144-158.

Gaffney, R., &Kirkby, R. (2018). Machine Learning the Consumption Function. EEA-ESEM Cologne 2018 Conference.

https://editorialexpress.com/conference/EEAESEM2018/program/EEAESEM2018

(Erişim Tarihi: 15.07.2020).

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (2005). Robust Statistics: The Approach Based on Influence Functions. John Wiley&Sons, Inc., New York.

Kolmogorov, A. (1933). Sulla Determinazione Empirica di una Legge di Distribuzione. *G. Ist. Ital. Attuari,* 4, 83-91.

Mian, A., Rao, K., & Amir, S. (2013). Household Balance Sheets, Consumption, and the Economic Slump. *The Quarterly Journal of Economics*, 148, 1687–1726.

Obrizan, M., Torosyan, K., & Pignatti, R. (2019). Tobacco Spending in Georgia: Machine Learning Approach. ICDSIAI 2018: Recent Developments in Data Science and Intelligent Analysis of Information, 103-114.

Önder, K., &Turgut, H. (2018). Examination of the Factors Affecting Household Rental Housing Demand Through Data Mining: The Case of Turkey. *Eskişehir Osmangazi Üniversitesi İİBF Dergisi, 13*(2), 227-238.

Pedregosa, F. (2016). Hyperparameter Optimization with Approximate Gradient. 33rd ICML, New York, 2016,(Editör. M. F. Balcan and K. Q. Weinberger). Proceedings of Machine Learning Research, 48: 737-746.

Rao, C. R. (1973). Linear Statistical Inference and its Applications.2. Basım, John Wiley & Sons, Inc., Canada.

Sec, R., &Zemcik, P. (2007). "The Impact Of Mortgages, House Prices And Rents On Household Consumption In The Czech Republic", CERGE-EI Discussion Paper, 2007–2185

Selim, S., & Demirkıran, E. (2020). Türkiye'de Hanehalkı Gıda Harcamalarını Etkileyen Sosyo-Ekonomik Faktörler: Karşılaştırmalı Bir Analiz. *Hacettepe Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 38*(2), 297-321.

Shapiro S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika, 52*(3/4), 591-611.

Shi, P., & Tsai, C. L. (2002). Regression Model Selection a Residual Likelihood Approach. *J. R. Statist. Soc. B,* 64, 237-252.

Showers, V. E., & Shotick, J. A. (1994). The Effects of Household Characteristics on Demand for Insurance: A Tobit Analysis. *The Journal of Risk and Insurance, 61*(3), 492-502.

Smirnov, N. (1948). Table for Estimating the Goodness of Fit of Empirical Distributions. *Annals of Mathematical Statistics, 19*(2), 279-281.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, 58, 267-288.

TUİK. (2018). Hanehalkı Bütçe İstatistikleri Mikro Veri Seti, 2018, Metaveri, Amaç. İstanbul.

Wang, H,, LI, G., & JIANG, G. (2007). Robust Regression Shrinkage and Consistent Variable Selection Through the LAD-Lasso. *Journal of Business & Economic Statistics*, 25, 347-355.

Varlamova, J., &Larıonova, N. (2015). Macroeconomic and Demographic Determinants of Household Expenditures in OECD Countries. *Procedia Economics and Finance*, 24, 727-733.

Ylvisaker, D. (1977). Test Resistance. *Journal of the American Statistical Association, 72*(359), 551-556.