***ISSN:* 2536-4383**

Araştırma Makalesi

Hacettepe Üniversitesi İstatistik Bölümü Mezunlarının Mezuniyet Sürelerini Etkileyen Faktörlerin Lokal Regresyon (Loess) ile İncelenmesi

Durdu Karasoy [*a], ***Ecem Kumandaş*** [a], ***Nihal Ata Tutkun*** [a]

[a] *Hacettepe Üniversitesi, Fen Fakültesi, İstatistik Bölümü, 06800 Beytepe, Ankara, Türkiye.*

Öz

Loess, "Locally Estimated Scatterplot Smoothing" ve "Local Regression" ifadelerinin kısaltılmasıdır ve veri analizi için geleneksel en küçük kareler yönteminin genelleştirilmesi sürecidir. Loess, bağımlı ve bağımsız değişkenler arasındaki ilişkinin önsel bir özelliği sağlamasını gerektirmediği için parametrik olmayan bir yöntemdir. Çoğunlukla saçılım grafiğini düzleştirmek için kullanılmasına rağmen, çok değişkenli verilere de kolaylıkla genelleştirilebilir. Ayrıca, güven aralıkları ve diğer istatistiksel testler için de çıkarımsal işlemler vardır. Bu nedenlerden dolayı Loess, verileri incelemek için faydalı bir araçtır. Loess, bağımlı değişken ve bir ya da daha çok bağımsız değişkenler arasındaki ilişkinin grafiksel bir özetini sağlar.

Bu çalışmada, loess yönteminin uygulanabilirliğini göstermek için Hacettepe Üniversitesi İstatistik Bölümü'nün kuruluş yılı olan 1967 yılından 2015 yılına kadar mezun olanların mezun olma süreleri ve bölümü tercih sıraları ile akademik ortalamaları ele alınmıştır. Loess yöntemi sonucunda akademik ortalamanın mezuniyet süresini etkilediği, tercih sırasının ise etkilemediği görülmüştür. Aykırı değerlerin etkisini dikkate alan ve daha sağlam olan lowess yönteminin ise verilere daha uyumlu olduğu sonucuna varılmıştır.

Anahtar Kelimeler: Loess yöntemi; lowess yöntemi, düzgünleştirme parametresi, parametrik olmayan regresyon

Analyzing Factors Affecting the Graduation Time of Hacettepe University Department of Statistics Students Using Local Regression (Loess)

Abstract

The term "Loess" is an acronym for "Locally Estimated Scatterplot Smoothing" and "local regression" and the entire procedure is a fairly direct generalization of traditional least-squares method for data analysis. Loess is nonparametric in the sense that the fitting technique does not require an a priori specification of the relationship between the dependent and independent variables. Although it is used most frequently as a scatterplot smoother, loess can be generalized very easily to multivariate data; there are also inferential procedures for confidence intervals and other statistical tests. For all of these reasons, loess is a useful tool for data exploration. Loess provides a graphical summary of the relationship between a dependent variable and one or more independent variables.

With this study, it was aimed to examine the graduation time of students from the department of Statistics at Hacettepe University and determine the factors that affect the graduation time by loess method. Therefore, the data of students who graduate from the department from the foundation of the department to 2015 was used. It was seen with loess method that graduation time is effected by academic average, whereas it isn't effected by the order of preference. It was concluded that lowess method which considers the impact of outliers and is more robust is more appropriate to data.

Key Words: Loess method, lowess method, smoothing parameter, nonparametric regression

[*] Corresponding author
e-mail: durdu@hacettepe.edu.tr

Received: 18.05.2016
Accepted: 13.07.2016

Giriş

Regresyon analizi, regresyon fonksiyonu hakkında istatistiksel çıkarımda bulunan bir analizdir ve temelde iki değişken arasındaki ilişki yapısının incelenmesinde kullanılır [1, 2]. Parametrik ve parametrik olmayan regresyon teknikleri, regresyon analizine iki farklı açıdan yaklaşır. Parametrik regresyon güçlü varsayımlara sahipken, parametrik olmayan regresyon ise bu varsayımları gerektirmez [2, 3]. Bununla birlikte parametrik olmayan tahmin yöntemleri, parametrik model geçerli olduğunda, parametrik tahmin yöntemlerine göre daha az etkindirler. Parametrik regresyon ve parametrik olmayan regresyon yöntemleri, regresyon analizi için her ne kadar farklı yaklaşımlar olarak kabul edilseler de, bu durum bir yöntemin diğerini dışlayacağı anlamına gelmez. Parametrik olmayan regresyon parametrik regresyonun önerdiği modelin geçerliliğini doğrulamak için kullanılabilir ya da tam tersi, veriye uygun model parametrik olmayan model tarafından yapılan tahmine göre kurulabilir [1, 2]. Regresyon analizi; herhangi bir değişkenin (bağımlı değişken) bir veya birden fazla değişken ile (bağımsız – açıklayıcı değişken) arasındaki ilişkinin matematiksel bir fonksiyon biçiminde yazılmasıdır. Elde edilen bu fonksiyona ise regresyon denklemi adı verilmektedir [2].

Birçok çalışmada, bir bağımlı, bir ya da birden çok bağımsız değişken durumunda doğrudan regresyon çözümlemesi uygulanmakta, uygunluğuna ise dikkat edilmemektedir. Oysa ki, doğrusal ilişki söz konusu olmadığında parametrik olmayan yöntemler kullanılmalıdır.

Bu çalışmada, parametrik olmayan regresyon analizi ile ilgili genel bilgiler verilmiş ve parametrik olmayan regresyon analizinde tahmin yöntemlerine değinilmiştir. Lokal regresyon hakkında bilgiler verilmiş, loess ve lowess yöntemler tanıtılmıştır. Loess tahmin edicisinin nerelerde kullanıldığı, bant genişliği veya düzeltme parametresinin her yerel polinomiyal için nasıl kullanıldığı anlatılmıştır. Loess yönteminin özellikleri, avantajları ve dezavantajları hakkında bilgiler verilmiştir. Uygulama bölümünde ise gerçek bir veri kümesi kullanılarak parametrik olmayan regresyon modelinde yerel polinomiyal düzleştiricisi için loess yöntemi kullanılmış ve istatiksel sonuçlar elde edilmiştir.

Bu çalışmanın amacı, parametrik regresyon modelinin uygulanamayacağı durumda kullanılabilen loess yöntemini tanıtmak, avantajlarını ortaya koymak ve Hacettepe Üniversitesi İstatistik Bölümü'nün kurulduğu 1967 yılından 2015 yılına kadar mezun olan 3182 kişinin verilerini kullanarak mezuniyet süresini etkileyen faktörleri doğrusal olmayan ilişkinin parametrik olmayan regresyon yöntemi olan loess yöntemi ile incelemek olmuştur.

Materyal ve Yöntem

Loess:

Loess, değişkenler arasındaki ilişkileri belirlemek için grafiksel bir araçtır, verilere düzgün bir eğri uydurmak için basit bir stratejidir [4]. Loess'da istenen saçılım grafiğinin düzgünleştirilmesidir [5]. "Loess" terimi "lokal regresyon" ifadesinin kısaltmasıdır. Loess, bağımlı ve bağımsız değişkenler arasındaki ilişkileri, önsel bir bilgi gerektirmeden belirleyen bir teknik olduğu için parametrik olmayan bir tekniktir ve güven aralıkları ile diğer istatistiksel testler için de sonuçlar alınabilir [4, 6].

Loess, doğrusal ve doğrusal olmayan en küçük kareler regresyonu gibi klasik yöntemler üzerine inşa edilen modern modelleme yöntemlerinden biridir. Loess, Cleveland (1979) tarafından önerilmiş, Cleveland ve Devlin (1988), Cleveland, Devlin ve Grosse (1988), Cleveland ve Grosse (1991) tarafından geliştirilmiştir [6, 7, 8, 9]. Lokal olarak ağırlıklandırılmış polinomiyal regresyon olarak da bilinir. Loess uyumu, n veri noktalarının her biri için hesaplanan regresyon fonksiyon değerlerinden sonra tamamlanır. Bu yöntemin detaylarından çoğu (polinomiyal modelin derecesi, ağırlıklar gibi) esnektir. Loess'daki her bir ağırlıklandırılmış en küçük kareler uyumu için kullanılan verilerin alt kümeleri en yakın komşuluk algoritmasıyla belirlenir [10].

Kullanıcının belirlediği, bant genişliği ya da düzleştirme parametresi olarak ifade edilen girdi, her bir lokal polinomiyal uyumu için kullanılır. Düzleştirme parametresi (d+1)/n ile 1 arasında bir sayıdır. Burada d, lokal polinomiyalin derecesini gösterir. Düzleştirme parametresinin değeri her bir uyumda kullanılan verinin oranıdır [11].

Loess'un aykırı değerlere hassasiyetini azaltmak için loess'un sağlam (robust) versiyonu kullanılabilir [11].

Loess, bir bağımlı değişkenle bir ya da daha çok bağımsız değişken arasındaki ilişkinin grafiksel bir özetini sağlar. Bu sürecin ayırt edici bir özelliği, verilerin kendi kendine konuşmasıdır. Loess, veriler arasındaki sistematik yapı için çok faydalı bir araçtır. Böylece bu teknik, teorileri geliştirmek için araştırmacılara yardım eder [4].

Loess yani polinomiyal regresyon, genel doğrusal regresyon modelinin özel bir halidir. Regresyon fonksiyonunu eğrisel hale getirmek için, bağımsız değişken(ler)in karesini ve yüksek dereceden değerlerini içerir [2].

Lokal olarak ağırlıklı regresyon, düzgünleştirilmiş noktaları bozan olağandışı noktalardan korunma amacı ile sağlam hale getirilebilir. Bu amaçla sürece, iteratif ağırlıklı en küçük kareler yöntemi adapte edilir. Böylece, sağlam lokal olarak ağırlıklı regresyon, düzgünleştirme ile sağlam tahminin bir kombinasyonu biçiminde elde edilir. Lowess yönteminde; lokal olarak kullanılacak olan polinomiyalin derecesi, ağırlık (kernel) fonksiyonunun ne olacağı, modeli sağlam hale getirmek için kaç iterasyon yapılması gerektiği ve düzgünleştirmede kullanılacak gözlem oranı (span) konuları karar verilmesi gereken kritik noktalardır [7].

Saçılım Grafiği:

İki boyutlu saçılım grafiği, iki değişken için basit bir grafiksel gösterim yöntemidir. Aynı zamanda, saçılım grafiği, çok değişkenli verilerin daha kompleks grafiksel gösterimleri için bir blok yapısı verir. Saçılım grafikleri, değişkenler arasındaki ilişkileri ya da fonksiyonel bağımlılıkları değerlendirmede görsellik sağlar. Noktalar, saçılım grafiğinin bir ölçek ekseni üzerinde farklı koordinatlara sahipse, ayrıca diğer ölçek ekseni üzerinde sistematik olarak farklı koordinantlar sergilemeye eğilimli ise fonksiyonel bağımlılık olduğu ifade edilir. İki değişken birbiriyle ilişkili ise, yani fonksiyonel bağımlı iseler bu durumda grafikteki noktalar, tam olarak grafik alanı boyunca tekdüze dağılmayacaklardır. İki değişken arasında ilişki olmadığında yatay bir çizgi elde edilecektir [4].

Loess eğrisi:

n gözlemli, X ve Y değişkenli bir veri seti olduğunda, bu veriler, X değişkeni yatay eksende, Y değişkeni dikey eksende olacak biçimde iki değişkenli bir saçılım grafiği elde edilir. Grafikteki noktalar (x_i, y_i) (i=1, 2, …, n) ikililerinden oluşur [4].

Loess süreci, m lokasyon serilerinin ya da v_j (j=1, 2, …, m) değerlendirme noktalarının seçimi ile başlar. Bu değerlendirme noktaları, X'in dağılım genişliğinin eşit aralıklarına karşılık gelir. Y değişkeninin koşullu dağılımlarındaki değişkenlik hakkında yeterli detayı sağladığı sürece; değerlendirme noktalarının tam sayısı nispeten önemsizdir. Pratikte, genellikle loess eğrisini uydurmak için kullanılan paket program tarafından belirlenir.

Daha sonra, Loess v_j'nin her birinde m ağırlıklandırılmış regresyon analizi serisi uygular. Bu regresyonlar ''lokal'' olarak ifade edilir, çünkü her biri sadece, saçılım grafiğinin yatay ekseni boyunca değerlendirilen noktalara en yakın düşen gözlemlerin altkümelerini kullanır. Araştırmacılar, loess parametresi olarak adlandırılan α'yı kullanarak her bir alt kümede içerilen toplam verinin oranını belirler. Lokal regresyonlar ya doğrusal ya da karesel eşitlikleri kullanabilirler. Loess λ parametresi kullanılarak fonksiyonel form belirlenir. Diğer bir durumda, her bir lokal regresyonda içerilen gözlemler, X ekseni boyunca değerlendirilen noktalardan onların uzaklığına göre ters olarak ağırlıklandırılır. Bu ağırlıklar, lokal regresyon içerisinde daha uzak düşen gözlemlerden, lokal regresyon çizgisinin (ya da eğrisinin) üzerinde daha çok etkili olacak olan v_j'ye yakın olan gözlemleri garantilerler. Lokal regresyonlar ayrıca, olağan dışı veri noktalarının etkisini azaltmak için sağlam tahmin edici sürecini de içerir.

Her bir lokal regresyondan katsayılar, tahmin edilen (predicted) ya da uydurulan (fitted) değeri tahmin etmede kullanılır. Değerlendirilen nokta için $\hat{g}(v_j)$ ile belirlenir. Lokal regresyonların tümü tamamlandıktan sonra, m farklı sıralı ikililer (v_j, $\hat{g}(v_j)$), n veri noktaları üzerinden üst üste saçılım grafiğinde çizdirilir.

Sonuç olarak, bitişik, yakın komşuluklu noktalar, yani ardışık v_js için (v_j, $\hat{g}(v_j)$), ilişkilidir. Değerlendirilen noktalar yatay eksen boyunca her biri diğerine yakın olarak yer alır. Bu ilişkili çizgi parçaları, veri noktaları boyunca geçen düzgün eğri olarak görünür [4].

Loess süreci, saçılım grafiğinin yatay ölçek ekseni boyunca bazı zamanlar "dikey sürgülü pencere" (vertical sliding window) olarak ifade edilir. Pencere durur ve m farklı v_js'nin her birinde ayrı regresyon eşitliğini ağırlıklandırılmış en küçük kareleri kullanarak tahmin eder. Regresyonlar sadece pencere içine düşen veri noktalarını içerdiğinden, tahmin edilen eğriler (ve bu durumda fit edilen değerler), verinin şeklini (dış hatlarını) izlemeyi değiştirebilir [4].

Loess eğrisi için parametrelerin bulunması:

Loess süreci parametrik değildir. Bu nedenle, final eğrinin fonksiyonel bir formu belirlenmemiştir. Fakat bazı parametreler vardır. Bu parametreler, loess eğrisinin gerçekten deneysel veri noktalarının merkezinden geçmesini garantilemek için uyarlanan sürece öncelik (prior) sağlamalıdırlar. Bu parametreler için seçilen değerler, kararlar oldukça basit olacak biçimde subjektif olarak belirlenir [4].

Düzleştirme parametresi, α:

Loess'da kayan pencere genişliğini (width of the sliding window) α parametresi belirler. Daha özel olarak α, her bir lokal regresyonda kullanılan gözlemlerin oranını verir. Bu parametre 0 ile 1 arasında bir değer olarak belirlenir. Örneğin α=0,65 olarak alınırsa, bu durumda lokal regresyonun her biri, toplam veri noktalarının %65'ini birleştiren eğri sürecini kullanır. Bu parametrenin büyük değerlerinde eğri düzleşir.

Loess eğrisi, veriler içerisinde X değerlerinin dağılım genişliği boyunca uzanan bir dizi olarak düşünülebilir. α değeri, bu dizinin "gevşekliğini" (slackness) kontrol eder. Daha büyük değerler daha sıkı çekerek, daha düzgün eğri üretir. Bu nedenle α, bazı zamanlar tansiyon, germe (tension) parametresi olarak da ifade edilir ve bant genişliği de denir.

α'nın küçük değerlerinde pencere genişliği aşırı derecede dardır ve bu durumda lokal regresyon "gürültü" (noise) varyasyonuna oldukça duyarlıdır. Bu nedenle eğriler çok faydalı olmaz, tercih edilmez. α'nın büyük değerlerinde de (0,75; 1) eğriler çok düz olur, fakat onlar da noktaların merkezinden geçmede başarısız olurlar. α'nın ara bir değeri ise α'nın küçük değerlerindeki durumla (aşırı uyum), α'nın büyük değerlerindeki durum (uyum kaybı) arasında uzlaşma sağlar. Genel amaç, verilerdeki var olan önemli yapının tümünü yakalayabilen loess eğrisini mümkün olduğunca düz elde etmektir [4].

Olması gerekenden daha geniş bir bant genişliği kullanıldığında tahminde daha küçük bir varyans elde edilir; fakat gerçek yoğunluk aşırı düzgünleştirilmiş olabileceği için yanlı bir tahmin elde edilir. Daha dar bir bant genişliği kullanılırsa bu kez de yanlılık miktarı azalırken, tahminin varyansı artabilir. Literatürde yanlılık ve varyans arasındaki bu dengeyi kurmak amacı ile çeşitli kriterler kullanılmaktadır. Karesel kaybı (quadratic loss) ölçen "Hata Kareler Ortalaması (Mean Squared Error-MSE)" bu kriterlerden biridir [2].

Loess polinomiyal derecesi, λ:

λ parametresi, loess sürecindeki polinomiyalin derecesini belirler. Eğer λ=1 ise, pencerelerin her biri içerisinde doğrusal eşitlikler uyarlanır. λ=2 olduğunda, karesel eşitlikler kullanılır [4].

λ parametresinin belirlenmesi pratikte, oldukça kolaydır. Karar saçılım grafiğinden verilebilir. Eğer noktalar bulutu monoton bir görüntüdeyse (artan ya da azalan) lokal olarak doğrusal uyum için λ, 1 olarak seçilir. Eğer veriler, lokal minima ve/veya maksima ile monoton olmayan bir görüntüdeyse lokal karesel eşitlikler için λ, 2 olarak seçilir [4].

Eğer X ve Y bir monotonik ilişki sergiliyorsa, bu durumda lokal uyarlanan pencerelerdeki nokta bulutları daima aynı genel görüntüye yönelirler. Saçılım grafiğindeki monotonik olmayan görüntüler durumunda ise doğrusal ilişkilerle ifade edilemez [4].

Polinomiyalin derecesi yanı etkiler. Polinomiyalin derecesi arttıkça yan azalır. Genellikle düşük dereceden polinomiyaller seçmek daha etkilidir. En çok kullanılan polinomiyaller lokal doğrusal ve lokal kareseldir. Lokal doğrusal daha çok yana sahiptir fakat özellikle sınırlarda varyans düşüktür. Kübik ve yüksek dereceden polinomiyaller daha çok uyum geliştirmemeye eğilimlidirler [12].

Ağırlık fonksiyonunun seçimi yan-varyans üzerinde daha az etkilidir. Ancak, uydurulan regresyon eğrisinin görsel kalitesini etkiler. Parametrik olmayan regresyon modelleri için en yaygın kullanılan ağırlık fonksiyonu tricube ağırlık fonksiyonudur [12].

Parametrik olmayan modeller için serbestlik derecesi konusu doğrusal modellerde olduğu gibi kolay değildir. Çünkü tahmin edilen parametreler yoktur. Parametrik olmayan model için sertbestlik derecesi, parametrik modeldeki parametrelerin sayısının genellemesidir. Yaklaşık serbestlik derecesi kullanılarak aynı veri kümesine uygulanan farklı tahminleri karşılaştırmak için F testi uygulanabilir (uygulanan polinomiyallerin farklı düzeylerini karşılaştırmak ya da doğrusal model ile düzleştirilmiş modeli karşılaştırmak gibi). Ayrıca, sertbestlik derecesi, uyarlanan eğri etrafındaki güven sınırlarını belirlemek için de gereklidir [12].

Loess'da katsayı (parametre) tahminleri yoktur. Bu nedenle parametrik olmayan olarak ifade edilir. X ve Y arasındaki ilişki grafikle gösterilebilir. Eğri uydurmayla ilgilenilir. Tahminler ve güven aralıkları hesaplanabilir fakat onlar belirli tahminlerden ziyade tamamlanan eğriye göredir. Diğer bir ifadeyle, kitleyi ifade eden eğrinin nasıl en iyi bir biçimde tahmin edilebileceğine odaklanılır [12].

Loess artık grafikleri:

Loess'da artıklar, düzleştirme eğrisinin verilerdeki ilginç yapının tümünü uygun olarak birleştirip birleştirmediğini belirlemek için kullanılan bir araçtır. Artıklar sistematik örüntü için irdelenmektedir [4].

Loess için eşitlik,

$$y_i = \hat{g}(x_i) + e_i \qquad (1)$$

biçiminde verilir. Bu durumda loess artıkları Y değişkeninin gözlenen değerleri ile X değişkenine göre karşılık gelen uyum değerlerinin farkı olarak Eşitlik 2'deki gibi tanımlanır:

$$e_i = y_i - \hat{g}(x_i) \qquad (2)$$

Eşitlik 2, regresyon çözümlemesindeki artıkların hesaplanması için kullanılan formüle çok benzerdir. Fakat, önemli bir fark vardır. Loess eğrisini bulmada kullanılan m değerlendirme noktaları hayali (imaginary) değerlerdir. Bu hayali değerler, X bağımsız değişkenin gözlenen değerlerinden genellikle farklıdır. Bu nedenle, deneysel gözlemler için uyarlanan değerler yani $\hat{g}(x_i)$'ler tipik olarak eşit aralıklı değerlendirme noktalarının en yakın iki oluşumları arasında interpolasyon ile elde edilir. Loess artıkları elde edilince, ya orijinal X değişkeninin değerlerine karşı ya da daha genel olarak karşılık gelen uyarlanan değerlere karşı çizdirilir. Bu durumda, loess eğrisi, artık çizimindeki noktalara uyarlanır. Loess düzleştiricisinin bu yeni uygulaması, artık çizimindeki dikey eksendeki sıfır değerinde yer alan düz bir çizgi üretir. Artıklar arasında herhangi bir örüntü olmamalıdır [4].

Loess düzleştrime eğrileri için uyum iyiliği:

Loess için, R^2 değerine benzer bir özet uyum istatistiği, loess uyum değerlerinin kareler toplamının bağımlı değişkene ilişkin kareler toplamına oranından aşağıdaki eşitlik ile hesaplanabilir:

$$R^2_{loess} = \frac{\sum_i^n(\hat{y}_i-\bar{\hat{y}}_i)^2}{\sum_i^n(y_i-\bar{y})^2} \qquad (3)$$

Burada,

y_i ; bağımlı değişkenin gerçek değerini,

$\hat{y}_i$; bağımlı değişkenin loess ile elde edilen uyum değerini ($\hat{g}(x_i)$),

$\bar{y}$; bağımlı değişkenin ortalamasını,

$\bar{\hat{y}}_i$; loess ile elde edilen uyum değerlerinin ortalamasını,

n ; gözlem sayısını gösterir.

Düşük R^2_{loess} değeri, bağımlı değişkendeki toplam dağılımın bir kısmını (tümünü değil) düzleştirme eğrisinin özetleyebildiğini gösterir. R^2_{loess}, varyans açıklanma oranı olarak yorumlanmaz, çünkü loess uyum süreci, y'deki kareler toplamının bir bölümü değildir. Bu nedenle R^2_{loess} değerinin varyans oranını verdiğini söylemek uygun değildir. Ancak, R^2_{loess} daha kısıtlı olarak yorumlanabilir [4].

Loess ve istatistiksel sonuç:

"X ve Y arasında fonksiyonel bağımlılık yoktur" yokluk hipotezine karşı test için Eşitlik 4 kullanılabilir:

$$F = \frac{(KT_Y - AKT_{loess})/(sd_{loess} - 1)}{(AKT_{loess})/(n - sd_{loess})} \qquad (4)$$

Burada KT_Y, Y için kareler toplamını, AKT_{loess}, loess için artık kareler toplamını, n gözlem sayısını gösterir.

Loess için artık kareler toplamı doğrusal regresyondaki gibi hesaplanabilir. Ancak, loess eğrileriyle ilişkili serbestlik derecesi biraz yanıltıcıdır ve hesaplanması oldukça karmaşıktır. Serbestlik derecesi genellikle tam sayı değildir. Bu yüzden sezgisel olarak alınabilir ya da bazı hesaplama yöntemleri kullanılabilir. Genellikle de parametre sayısı serbestlik derecesi olarak alınır [4].

Aynı veriye uyarlanan iki loess uyumunu karşılaştırmak için varyans analizi kullanılabilir. Geleneksel doğrusal regresyondaki gibi bir test istatistiği elde edilebilir. Bu test istatistiği Eşitlik 5 ile verildiği gibidir:

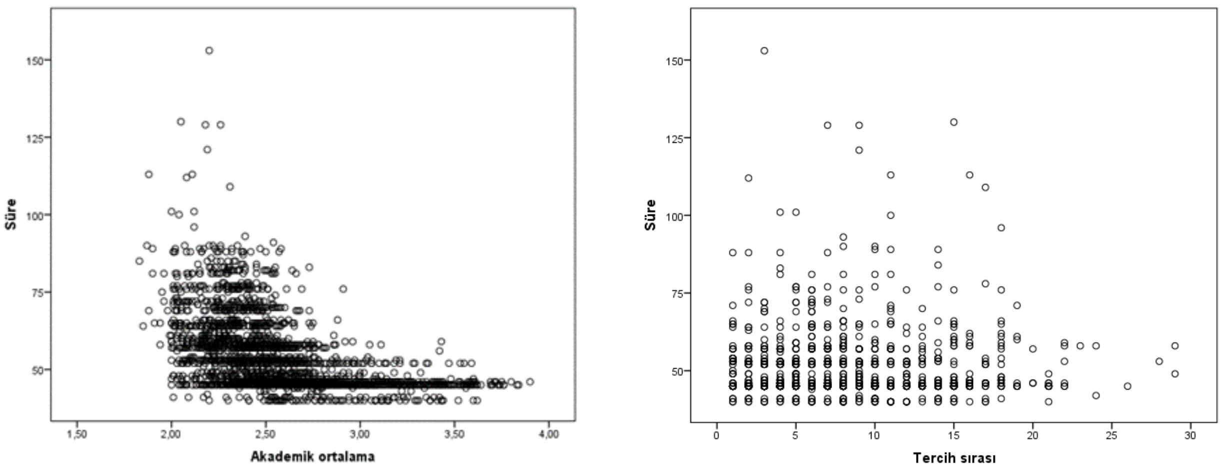
$$F = \frac{(AKT_{loess,1} - AKT_{loess,2})/(sd_{loess,2} - sd_{loess,1})}{(AKT_{loess,2})/(n - sd_{loess},2)} \qquad (5)$$

Burada $AKT_{loess,1}$ ve $AKT_{loess,2}$, iki eğrinin loess artıkları kareler toplamını, $sd_{loess,1}$ ve $sd_{loess,2}$ ise eğriler için serbestlik derecelerini, n gözlem sayısını gösterir. "İki model uyumunda bir farklılık yok" yokluk hipotezi altında test yapılır. Loess1 doğrusal model, loess2 karesel modelse ve yokluk hipotezi reddedilirse, loess eğrisi önemlidir, doğrusal modele göre önemli bir gelişme göstermektedir biçiminde yorumlanır [4].

Bulgular

Bu çalışmada, Hacettepe Üniversitesi İstatistik Bölümü'nün kurulduğu 1967 yılından 2015 yılına kadar mezun olan 3182 kişinin verileri kullanılarak bir uygulama yapılmıştır. Mezuniyet süresi (52,78 ay±0,21) ile tercih sırası (8±0,14) ve akademik ortalama (2,6±0,06) arasındaki ilişkiyi belirlemek için loess yöntemi uygulanmış ve yorumlanmıştır.

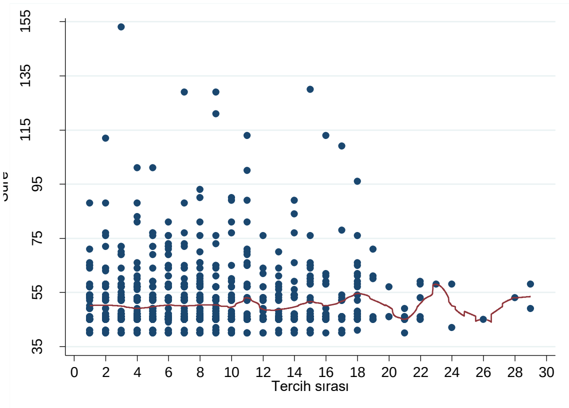
Akademik ortalama ve tercih sırası değişkenlerine ilişkin saçılım grafiği Şekil 1'de verilmiştir.



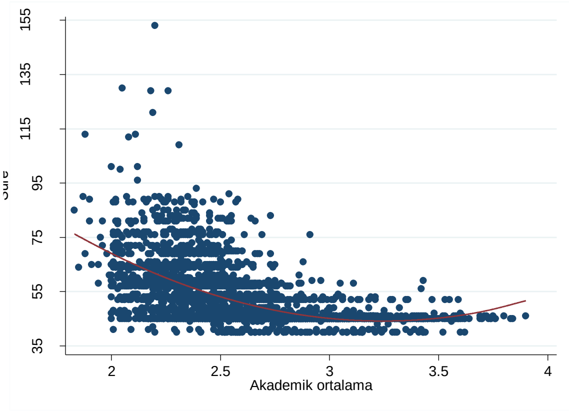
Şekil 1. *Akademik ortalama ve tercih sırası değişkenleri için saçılım grafiği*

Şekil 1 incelendiğinde mezuniyet süresi ile akademik ortalama ve mezuniyet süresi ile tercih sırası değişkenleri arasında doğrusal bir ilişki olmadığı görülmektedir. Bu durumda doğrusal regresyon çözümlemesini uygulamak doğru olmamaktadır. Doğrusal olmayan ilişkileri ortaya çıkaran loess yöntemini uygulamak daha doğru olmaktadır.

Her bir noktanın düzleştirilmesinde verilerin en yakın komşuluktaki %80'ninin kullanıldığını ifade eden bant genişliği 0,8 olarak alınarak tercih sırası ve süre için loess uygulandığında Şekil 2, akademik ortalama ve süre için loess uygulandığında ise Şekil 3 elde edilmektedir.



Şekil 2. *Tercih sırası ve süre için loess, düzleştirme parametresi = 0,8; derece = 2*



Şekil 3. *Akademik ortalama ve süre için loess, düzleştirme parametresi = 0,8; derece = 2*

Tercih sırası ve akademik ortalama değişkenlerinin her ikisi için düzleştirme parametresi bir büyük (0,8), bir küçük (0,3) ve bir de orta değer (0,5) alınarak çözümlemeler yapılmıştır. Ancak, tercih sırası mezuniyet süresini etkilemeyen değişken olarak bulunduğundan sadece 0,8 sonucu verilmiştir. Saçılım grafiklerinde, tercih sırası monoton artan ya da azalan bir görüntü vermediği için derecesi 2 olarak alınmış, akademik ortalama değişkeni için ise 1 ve 2 alınmış ve sonuçları karşılaştırılmıştır.

Tercih sırası ve akademik ortalama için loess sonuçları Tablo 1'de özetlenmiştir.

Tablo 1. *Tercih sırası ve akademik ortalama için loess sonuçları*

	Düzleştirme parametresi (bant genişliği)	**Derece**	**R^2_{loess}**	**F**	**HKO***
Tercih sırası	0,8	2	0,04	0,56	141,75
Akademik ortalama	0,8	2	0,62	444,02	175,86
	0,5	2	0,59	442,436	171,19
	0,3	2	0,58	436,26	170,2
	0,8	1	0,66	777,38	181,54
	0,5	1	0,56	652,67	165,83
	0,3	1	0,53	620,08	163,38

* HKO: Hata Kareler Ortalaması

Tablo 1 incelendiğinde, tercih sırası değişkeni için R_loess^2= 0,04 olarak elde edilmiştir. R_loess^2 değeri çok küçük olduğundan mezuniyet süresi değişkenindeki toplam dağılımı, tercih sırasına ilişkin düzleştirme eğrisinin özetleyemediği söylenebilir.

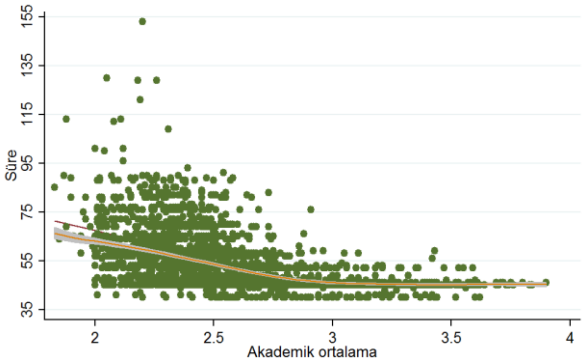
Tercih sırası ile mezuniyet süresi arasında fonksiyonel bağımlılık yoktur, yokluk hipotezi test edildiğinde F=0,56 olarak bulunduğundan ve bu değer 0,05 anlamlılık (F3,2944=2,6) düzeyinde önemli olmadığından tercih sırası ile mezuniyet süresi arasında fonksiyonel bir bağımlılık olmadığı %95 güvenle söylenebilir.

Akademik ortalama değişkeni için 0,8 bant genişliği ve 2. derece için R_loess^2= 0,62 olarak elde edilmiştir. R_loess^2 değerine göre mezuniyet süresi değişkenindeki toplam dağılımın önemli bir kısmını akademik ortalamaya ilişkin düzleştirme eğrisinin özetleyebildiği söylenebilir.

"Akademik ortalama ile mezuniyet süresi arasında fonksiyonel bağımlılık yoktur", yokluk hipotezi test edildiğinde F=444,02 olarak bulunduğundan ve bu değer 0,05 anlamlılık düzeyinde önemli (F3,2944=2,6) olduğundan akademik ortalama ile mezuniyet süresi arasında fonksiyonel bir bağımlılık olduğu %95 güvenle söylenebilir. Diğer değerler için de benzer yorumlar yapılabilir.

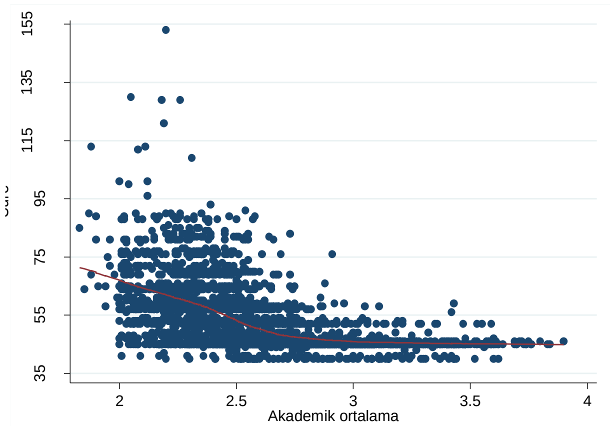
Tablo 1'den (HKO değerleri incelendiğinde) akademik ortalama için birinci dereceden loess'da ve ikinci dereceden loess'da 0,3 düzleştirme parametresi için elde edilen sonuçların daha iyi olduğu görüldüğünden aynı veriye uygulanan bu iki loess uyumunu karşılaştırmak için elde edilen F test istatistiği 82,01 olarak elde edildi ve iki loess uyumu arasındaki farklılığın anlamlı olduğu, birinci dereceden loess uyumunun veriler için daha uygun olduğu sonucuna varıldı.

Birinci dereceden ve 0,3 düzleştirme parametresiyle elde edilen loess uyumunun güven aralıklarını da içeren grafiği Şekil 4'te verilmiştir. Güven aralıklarının oldukça dar bir alanda olduğu görünmektedir.



Şekil 4. *Akademik ortalama ve süre için loess ve güven aralığı, düzleştirme parametresi = 0,3; derece = 1*

Aykırı değerlerin etkisini azaltmak için, düzgünleştirilmiş noktaları bozan olağan dışı noktalardan korunmak için, tricube ağırlık fonksiyonu kullanılarak düzleştirme parametresi 0,3 ve derecesi 1 alınarak daha sağlam bir yöntem olan lowess uygulandığında ise Şekil 5 elde edilmiştir.



Şekil 5. *Akademik ortalama ve süre için lowess, düzleştirme parametresi = 0,3; derece = 1*

Akademik ortalama değişkeni için R^2_{lowess}= 0,31 olarak elde edilmiştir. R^2_{lowess} değerine göre mezuniyet süresi değişkenindeki toplam dağılımın bir kısmını akademik ortalamaya ilişkin düzleştirme eğrisinin özetleyebildiği söylenebilir.

"Akademik ortalama ile mezuniyet süresi arasında fonksiyonel bağımlılık yoktur", yokluk hipotezi test edildiğinde F=654,66 olarak bulunduğundan ve bu değer 0,05 anlamlılık düzeyinde önemli ($F_{2,2945}$=3) olduğundan akademik ortalama ile mezuniyet süresi arasında fonksiyonel bir bağımlılık olduğu %95 güvenle söylenebilir.

Loess sonucunda elde edilen uyum ile lowess sonucunda elde edilen uyum arasında farklılığın anlamlılığı incelendiğinde F=2499 bulunduğundan lowess sonucu elde edilen uyumun diğerinden daha iyi olduğu sonucuna varılmıştır.

Sonuç

Regresyon analizi, regresyon fonksiyonu hakkında istatistiksel çıkarımda bulunan bir analizdir ve bir bağımlı, bir ya da birden fazla bağımsız değişken arasındaki ilişki yapısını incelemek için kullanılır. Ancak, regresyon analizinde sağlanması gereken varsayımları vardır.

Loess ise genel doğrusal regresyon modelinin özel bir halidir ve parametrik olmayan regresyon modeller ailesinin bir üyesidir. Loess verilere çok esnek bir yaklaşım sağlar. Saçılım grafiğini düzleştirmede kullanılır ve değişkenler arasındaki ilişkilerin fonksiyonel yapısını değerlendirmek için oldukça faydalı bir yöntemdir. Bu yöntemde veriler bir eşitlikle ifade edilemez ve parametreler için keyfi karar vermeyi gerektirir. Ancak yine de, parametrik olmayan düzleştiriciler, veriyi kendi kendine konuşturmada, geleneksel parametrik yöntemlerden daha iyidir. Bu durum özellikle doğrusal olmayan ilişki durumlarında önemlidir.

Literatürde, Cleveland (1979) sağlam lokal olarak ağırlıklandırılmış regresyon ve düzleştirme saçılım grafikleri hakkında, Cleveland, Devlin, Grosse (1988) lokal regresyon yöntemleri, özellikleri ve hesaplama algoritmaları hakkında, Cleveland ve Devlin (1988) lokal olarak ağırlıklandırılmış regresyon hakkında, Cleveland ve Grosse (1991) lokal regresyon için hesaplama yöntemleri hakkında, Jacoby (2000) ve Jacoby (2005) değişkenler arasındaki ilişkileri belirlemek için parametrik olmayan grafiksel bir araç olan loess hakkında, Tezcan (2011) parametrik olmayan regresyon analizi hakkında bilgiler vermişlerdir [2, 4, 6, 7, 8, 9, 12].

Bu çalışmada ise loess ve lowess yöntemleri hakkında bilgiler verilmiş ve bu iki yöntem kullanılarak, Hacettepe Üniversitesi İstatistik Bölümünün kurulduğu günden bu yana mezun olanların mezuniyet süresi üzerinde tercih sırası ile akademik ortalamanın etkisinin olup olmadığı incelenmiştir. Mezuniyet süresi üzerinde tercih sıralamasının bir önemi olmadığı, akademik ortalamanın ise önemli olduğu sonucuna varılmıştır. 0,3 değişim parametresiyle birinci dereceden lowess düzleştirmesinin verilere en uyumlu sonucu verdiği görülmüştür.

Bu çalışma, Hacettepe Üniversitesi İstatistik Bölümü'nden mezun olacak öğrencilerin mezun olmaları gereken sürede mezun olabilmeleri için akademik ortalamalarını yükseltmelerinin gerekliliğini, bunun için önlemlerin alınmasını, başarıyı arttırmak için öğrencilerin teşvik edilmesini göstermektedir.

Kaynaklar

[1] Eubank, R. L., 1999. Nonparametric regression and smoothing spline, Marcel Dekker Inc., New York.

[2] Tezcan N., 2011. Parametrik Olmayan Regresyon Analizi, Atatürk Ü. İİBF Dergisi, 10. Ekonometri ve İstatistik Sempozyumu Özel Sayısı, 341.

[3] Pehlivan N.Y., Apaydın A., 2005. Bulanık k-En Yakın Komşuluk Tahmin Edicisi ve Bulanık Radyal Tabanlı Fonksiyon Ağları , S.Ü. Fen Ed Fak Fen Derg., 26, 19- 32.

[4] Jacoby W.G., 2000. Loess: A nonparametric, graphical tool for depicting relationships between variables, Electoral Studies, 19, 577-613.

[5] Kara P., 2013. Loess; https://prezi.com/ankknumpnnu9/loess/ [Erişim 16.04.2016].

[6] Cleveland W.S., Devlin S.J., Grosse E., 1988. Regression by local fitting: Methods, properties, and computational algorithms, Journal of Econometrics, 37(1), 87–114.

[7] Cleveland W.S., 1979. Robust locally weighted regression and smoothing scatterplots, Journal of the American Statistical Association, 74, 829.

[8] Cleveland W.S., Devlin S.J., 1988. Locally weighted regression: An approach to regression analysis by local fitting, Journal of the American Statistical Association, 83(403), 596-610.

[9] Cleveland W.S., Grosse, E., 1991. Computational methods for local regression, Statistics and Computing, 1, 47–62.

[10] The Loess Procesure, SAS/STAT ® User's Guide, Version 8, Cary, NC: SAS Institute Inc., 1999; http://www.math.wpi.edu/saspdf/stat/chap38.pdf [Erişim 10.03.2016].

[11] Engineering Statistics Handbook;

http://www.itl.nist.gov/div898/handbook/pmd/section1/pmd14.htm [Erişim 16.03.2016].

[12] Jacoby B., 2005. Regression III: Advanced Methods, http://polisci.msu.edu/jacoby/icpsr/regress3; http://polisci.msu.edu/jacoby/icpsr/regress3/lectures/week4/15.Loess.pdf [Erişim 10.03.2016].