BÖLÜM 1

1.GİRİŞ

İstatistik biliminin temel uğraşılarından birisi de bir değişkenin davranışının tahminlenmesidir. Davranışı tahminlenecek olan değişken, başka değişkenlerin fonksiyonu olarak ortaya çıkabilir. Bu fonksiyonun elde edilmesi veya tahminlenmesi, incelenmesi ve yorumlanması regresyon analizinin ilgi alanı kapsamındadır. Davranışı diğer değişkenlerden etkilenen değişkene bağımlı değişken veya yanıt (response) adı verilir. Bağımlı değişkenler bir şans değişkenidir. Bağımlı değişken değerini etkileyen ve/veya yönlendiren değişkenler veya fonksiyonlarına açıklayıcı değişkenler adı verilir. Bu değişkenlerin kendi aralarındaki doğrusal bağımlılık yapılarının matematiksel olarak önemsiz olması durumunda her bir değişkene bağımsız değişken adı verilir. Bir açıklayıcı değişken bağımsız değişkenlerin fonksiyonu olabildiği gibi, yalnızca bir bağımsız değişkenin kendisi de açıklayıcı değişken olabilir. Temel regresyon analizinde bağımsız değişkenlerin bilinen sabitler olduğu varsayılır.

Bağımsız değişkenlerin, bağımlı değişkeni etkileyen fonksiyonuna *REGRESYON MODELİ* adı verilir. Varsayılan modelin, açıklayıcı değişkenlerin çalışma aralığı içinde, doğru bir model olması durumunda bağımlı değişkenlerin, modelin etkisi bertaraf edildikten sonra kalan yalın (pure) davranışı bir şans değişkeni davranışıdır. Bu yaklaşım daha ilerideki konularda ele alınacak olan hata terimlerinin davranışlarının temelidir.

MODEL; alternatif durumların alternatif sonuçlarını gösteren tablo, grafik, fonksiyon, prototip vb. yapılardır. Bu kitapta kullanılan model veya regresyon modelleri fonksiyon tipinde olan modellerdir.

Bu kitapta incelenen *MODELLEME* kavramı ise ilgilenilen bir şans değişkeninin davranışını tanımlamak amacıyla matematiksel bir yaklaşımın geliştirilmesini ifade eder. Bağımsız değişkenleri $X_1,...,X_k$ bağımlı değişkeni ise Y temsil eder. Araştırmada birden fazla gözlem alındığından gözlemleri birbirinden ayırmak için bir alt indis kullanılır Y_i. Sonuç olarak bir regresyon modelinin genel hali;

$$Y_i=f(X_1,...,X_k)+\varepsilon_i \qquad i=1, 2,\ldots,n$$

şeklindedir. Burada ε hata terimi olarak adlandırılır. Hata teriminin davranışı, bağımlı değişkenin hiçbir etkileyici faktör bulunmadığındaki davranışı olmalıdır. Eğer bu duruma ulaşılmış ise kullanılan regresyon modeli ile gerçek yaşam modeli arasındaki farklılıklar istatistiksel olarak önemsizdir. Bu şartı sağlayan model doğru model olarak kabul edilir.

MODELLEMENİN AMACI, koşullar değiştiğinde bağımlı değişkenin ortalamasının $E(Y_i)$ nasıl değiştiğini tanımlamaktır.

PARAMETRELER, tüm modeller bağımsız değişkenlere ilave olarak bilinmeyen sabitler içerir. Bilinmeyen bu sabitler *parametreler* olarak adlandırılır ve modelin davranışını kontrol eder. Bir

regresyon parametresi bir değişim oranıdır diğer bir ifade ile bir açıklayıcı değişkene göre bağımsız değişkenin kısmî türevidir.

HATA TERİMİ VE İSTATİSTİKSEL MODEL, bağımlı değişkenin elde edilen her bir gözlemi Y_i nin, ana kütle ortalaması $E(Y_i)$ olan bir ana kütleden gelen şans değişkeni olduğu varsayılır. Bir gözlemin Y_i, kendi ana kütle ortalamasından $E(Y_i)$ sapması, *matematiksel modele* bir *hata terimi* eklenerek açıklanır

MODELİN BELİRLENMESİ İÇİN GEREKLİLİKLER, olasılıksal modeller bir ya da daha fazla rassal bileşen (hata bileşeni) içerir. Her bir hata bileşeninin ait olduğu bir hata kaynağı mevcuttur.

Modelin tam olarak ifade edilebilmesi için hata teriminin istatistiksel özelliklerinin tanımlanması gereklidir.

Bir modelin özellikleri, hesaplandığı verilerden çok verilerin alındığı X in sınırlarına bağımlıdır. Bu nedenle model sadece tanımlanan X bölgesi için geçerlidir.

KESTİRİM, açıklayıcı değişkenlerin veri setinde bulunmayan bir değeri için elde edilen $\hat{Y}_i$ değeridir. İnterpolasyon ya da eksterpolasyon gerektirir.

Bağımsız değişkenlerin veri setindeki bir değeri için elde edilen $\hat{Y}_i$ ise *uyumu yapılan* değerdir.

MODELİN DOĞRUSALLIĞI, kullanılan modeller genellikle *parametrelerine göre doğrusaldır*. Bir modelin doğrusal olduğundan veya olmadığından bahsedildiğinde, ifade edilen, parametrelerdeki doğrusallık veya doğrusalsızlıktır. Parametrelere göre doğrusallık, modeldeki tüm parametrelerin birinci dereceden olmasıdır. Diğer bir deyişle üstel durumda ya da bir diğer parametre ile çarpım halinde veya bölüm halinde bir parametrenin bulunmamasıdır. Modeldeki bir bağımsız değişkenin en büyük kuvvetinin değeri modelin derecesidir. Örneğin;

$$Y_i = \beta_0 + \beta_1 X_i + \beta_{11} X_i^2 + \varepsilon_i$$

modeli ikinci dereceden (en büyük X derecesi) doğrusal regresyon modelidir. Bir model özellikle doğrusal olmayan model olarak belirtilmedikçe parametreleri açısından doğrusal olduğu kabul edilecektir, doğrusal kelimesi genellikle ihmal edilmekte ya da unutulmaktadır. Modelin derecesi herhangi bir büyüklükte olabilir. β_{11} notasyonu polinom modellerde kullanılmaktadır, bu parametre X^2 değişkeni ile birlikte kullanıldığında, X değişkeni için de β_1 parametresi kullanılır.

DOĞRUSAL OLMAYAN MODELLER, daha gerçekçi modeller karmaşık olup *parametrelerine göre doğrusal değildir*. Doğrusal olmayan modeller iki sınıfa ayrılır: Doğrusal hale dönüştürülebilenler (bağımlı ya da bağımsız değişken üzerine dönüşüm ile), doğrusal hale dönüştürülemeyenler

TAHMİNLEME, genel olarak doğrusal model p adet parametreye β_i sahiptir. Bu parametreler, veri seti, $(X_{i1},\ldots,X_{ik,}\ Y_i)$, kullanılarak tahminlenir. Şans değişkeni Y deki değişkenlik her bir gözlenmiş veri çiftinin farklı değerler almasına neden olur. Gerçekte hata terimi ε u tahminlemek, her bir gözlem için farklı değerler aldığından, oldukça zordur. Buna karşın β_i parametreleri sabit değerler aldığı için veri seti kullanılarak tahminleri olan b_i değerleri elde edilebilir.

Regresyon fonksiyonu $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ şeklinde verildiğinde, X_i ve ε_i'nin Y üzerindeki etkilerinin ayrı ve eklenebilir olduğu kabul edilir. Bu modelde hata terimi ihmal edilmiş tüm değişkenlerin etkisini temsil edebilir. Fakat eğer X_i ve ε_i ilişkili ise, onların Y üzerinde bireysel etkilerini değerlendirmek mümkün değildir. Buna göre eğer X_i ve ε_i pozitif doğrusal ilişkili ise ε_i artarken X_i artacaktır veya ε_i azalırken X_i'de azalacaktır. Benzer olarak eğer X_i ve ε_i negatif ilişkili ise ε_i azalırken X_i artacaktır yada ε_i artarken X_i azalacaktır. Her iki durumda da X_i ve ε_i'nin Y üzerindeki etkisini ayırmak zordur. Bu nedenle klasik regresyon analizinde X_i değerlerinin sanş değişkeni olamadığı, bilinen sabitler olduğu kabul edilir.

Değişkenler arasındaki ilişkiler üç başlıkta ele alınıp sınıflandırılabilir. Regresyon analizinin ilgi alanı aşağıda belirtilen bu ilişkilerden yalnızca *yarı deterministik* ilişkiler ve *olasılıksal* ilişkileri kapsamaktadır.

DETERMİNİSTİK İLİŞKİLER, bağımsız değişkenlerin, bağımlı değişkeni hangi fonksiyonla belirlediğinin bilindiği ve bağımlı değişkenin şans değişkeni olmadığı durumundaki ilişkilerdir. Örneğin *a* liralık bir kapitalin dönemlik faiz oranı *i* olduğu durumda *n*-inci dönem sonundaki değeri

$$A_n = a(1+i)^n$$

fonksiyonu ile tanımlanır. Burada A_n bağımlı *a, i* ve *n* ise bağımsız ve açıklayıcı değişkenlerdir. *a, i* ve *n'* in aynı değerleri için daima aynı A_n değeri elde edilir. Bir başka deyişle A_n değişkeni şans değişkeni değildir. Bu kapsamdaki modeller regresyon analizi dışındadır.

YARI DETERMİNİSTİK İLİŞKİLER, yarı deterministik bir model, bağımlı değişkenin şans değişkeni olması durumudur. Böylece, modelin katsayılarının tahminleri de şans değişkeni olma özelliğini taşırlar. Örneğin; Cobb-Douglas üretim fonksiyonunun geçerli olduğu bir durum ele alınsın. *Y*' ler üretim miktarlarını, X_1 işgücü miktarını, X_2 sermaye miktarını, α ve β sırası ile işgücü ve sermayenin üretim miktarını etkilemede önemli olan katsayılarını, *k* ise bir sabiti göstersin.

$$Y_i = kX_1^{\alpha} X_2^{\beta} \varepsilon_i$$

Aynı firmada, X_1 ve X_2 değerlerinin değişmediği durumlarda dahi farklı *Y* değerleri söz konusu olacaktır. *Y* değerlerindeki bu değişkenlik saf hatanın bir göstergesidir. Bu durum *Y*' nin şans değişkeni olmasının bir sonucudur.

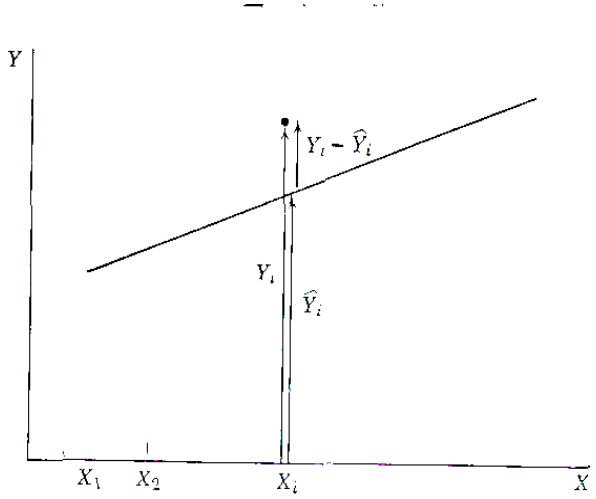
Esas olarak *Y*' nin toplam değişkenliği iki kısımda açıklanır. Birinci bileşen modelin açıkladığı değişkenlik, ikinci bileşen ise *Y*' nin saf değişkenliğidir. Birçok regresyon modelinin değerlendirilmesinde açıklanan değişkenlik yüzdesi modelin yeterliliği için bir araç olarak kullanılır. Ancak, bu her zaman yeterli olmayacaktır. Çünkü bağımsız değişken katsayısının arttırılması bu matematiksel sonucu hemen doğurabilir. Bu gibi durumlarda elde edilen yüksek açıklama yüzdeleri ise anlamsız veya önemsiz olabilirler. Amaç, kullanılan model ile *Y*' nin saf değişkenliğine ulaşmaktır.

OLASILIKSAL İLİŞKİLER, bu tür ilişkilerde Y bağımlı değişkenleri birer şans değişkenleridir. Aynı zamanda kullanılacak model ile ilgili bir ön bilgi, doğrulama mevcut değildir. Bu nedenle model bağımsız ve bağımlı değişkenlerin incelenmeleri sonucu tahmin edilmektedir. Bu tür durumların sakıncalı bazı yönleri bulunmakla birlikte, çok sık karşılaşılan durumlardır. Örneğin, bir şirketin satış miktarlarını ürün bazında etkileyen faktörler biliniyor olabilir. Bu faktörlere ait değişkenler tanımlanabilir. Açıklayıcı ve hatta bağımsız hale getirilebilirler. Bütün bunlara rağmen model (açıklama yapısı) karmaşık ve bilinmeyen konumundadır. Karmaşıklık etkileyen faktör sayısının çokluğunda ve hesaba katılmayan ancak etkileri küçük değişkenlerin zaman zaman ön plana çıkma biçimlerinden vb. nedenlerden oluşabilir.

Regresyon analizinde kullanılan modeller istatistiksel olarak anlamlı bulunsalar da, gerçek duruma uygunlukları uzmanlarınca değerlendirilmeli ve gerekli revizyonlar yapılarak, tahminlerin geçerliliğine süreklilik kazandırılmalıdır.

1.1 İYİ BİR UYUM İÇİN MÜMKÜN KRİTERLER

İyi bir uyum nedir? Bu sorunun cevabı, elbette ki *toplam hatayı minimum yapan uyum, iyi bir uyumdur* şeklindedir. Regresyon analizinde amaç, kurulan model ile elde edilen tahminlenmiş $\hat{Y}_i$ değerlerinin gözlenmiş Y_i değerlerini yeterince küçük bir hata ile temsil edebilme yeteneğine sahip olabilmesidir. Tipik bir hatanın grafiksel ifadesi Şekil 1.1' de gösterilmektedir. Hatanın gözlenmiş (modelden tahminlenmiş) değeri artık olarak adlandırılır ve gözlenmiş Y_i değerleri ile uyum yapılmış doğru arasındaki dikey uzaklık $\left(Y_i - \hat{Y}_i\right)$ olarak tanımlanır. Yukarıdaki ifadede verilen $\hat{Y}_i$„ gözlenmiş Y_i değerinin uyumu yapılmış değeri başka bir deyişle doğrunun ordinatıdır. Gözlenmiş Y_i değeri doğrunun üst kısmında olduğunda artık pozitif, alt kısmında olduğunda ise artık negatiftir. Gerçekte verilere uyumu sağlanan matematiksel bir modeldir. İyi uyumu model kavramı altında istatistiksel olarak açıklarsak; modeli, etki eden faktörlerin etkisi ile doğal varyasyonu birbirinden ayırma derecesi olarak kullanabiliriz. Aşağıda anlatılanlar doğru modelin tespit edildiği varsayımı altında iyi uyum kriterleridir.

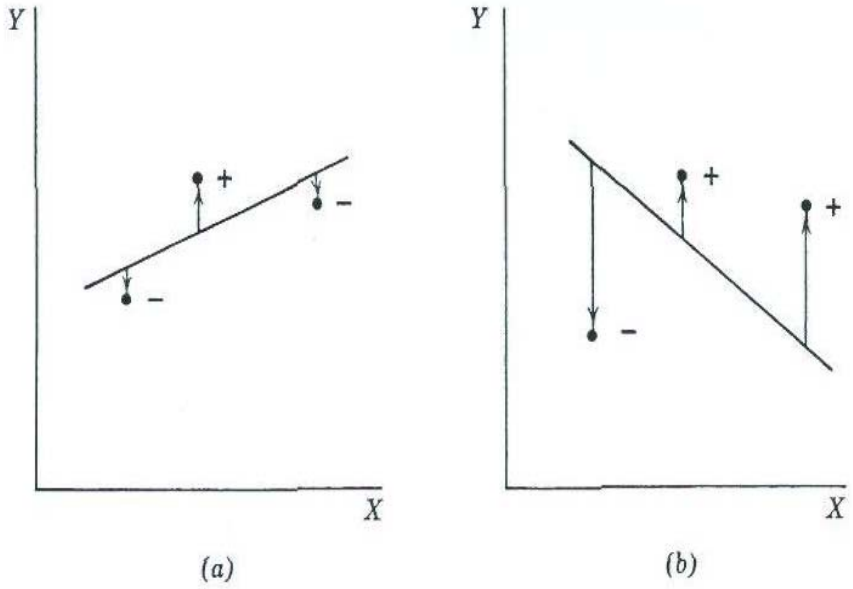


Şekil 1.1 *Bir doğruya uyumu yapılan noktalardaki hata* (*artık*)

Birinci kriteri değerlendirmek üzere, tüm hataların toplamını minimize eden uyumu yapılmış bir doğru dikkate alınacaktır. Hataların toplamı,

$$\sum\left(Y_i - \hat{Y}_i\right) \tag{1.1}$$

şeklinde ifade edilebilir. Bu kriteri kullanmak pek faydalı değildir. Bunun nedeni, aynı gözlemler kullanılarak uyumu yukarıdaki kritere göre yapılmış iki doğru üzerinde açıklanabilir. Şekil 1.2' de verilen bu iki doğrudan biri göreceli olarak iyi diğeri ise oldukça kötüdür. Her iki durumda da sorun işaretlerden kaynaklanmaktadır, pozitif ve negatif hataların toplamı sıfır değerini vermektedir. İyi ve kötü uyum arasındaki farkı ortaya koymadığı için bu kriter reddedilebilecektir.

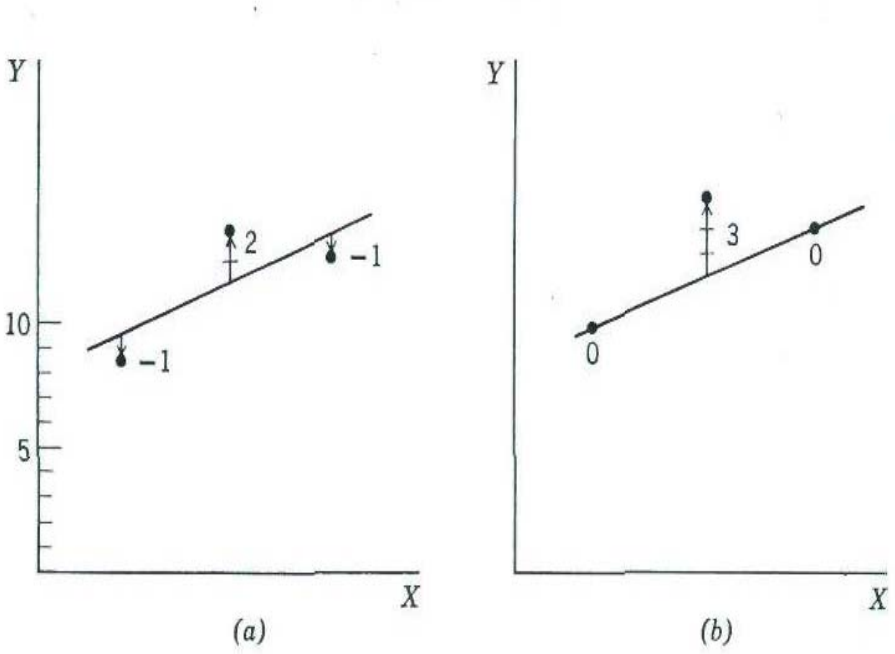


Şekil 1.2 $\sum\left(Y_i - \hat{Y}_i\right)$ *kriteri için aynı üç noktaya uyumu yapılan iki farklı doğru*

Yukarıda değinilen işaret problemini gidermenin iki yolu mevcuttur. Bunlardan birincisi hataların mutlak değerlerinin toplamını minimize etmektir.

$$\sum\left|Y_i - \hat{Y}_i\right| \tag{1.2}$$

Pozitif ve negatif değerler bu kriterde birbirlerini düzeltemeyeceklerdir. Böylece Şekil 1.2.*b*' deki gibi kötü uyumları engelleyecektir. Bununla birlikte bu kriterde de bir dezavantaj vardır. Şekil 1.3'de bu durum açıklanmaktadır. Verilen kritere göre (*b*)' deki uyum daha iyidir. Çünkü $\sum\left|Y_i - \hat{Y}_i\right| = 3$ olup 4' den küçüktür. (*b*)' deki doğru incelendiğinde bu doğrunun uç noktalar için en iyi doğru olduğu görülmektedir. Bununla birlikte ortadaki nokta dikkate alınmamaktadır. Bu nedenle probleme ortak bir çözüm getirmemektedir. (*a*)' daki doğru ise tüm noktaları dikkate aldığı için tercih edilebilecektir.



Şekil 1.3 $\sum\left|Y_i-\hat{Y}_i\right|$ *kriteri için aynı üç noktaya uyumu yapılan iki farklı doğru*.

İşaret problemini ortadan kaldıracak ikinci yöntem ise hataların kareler toplamını,

$$\sum\left(Y_i-\hat{Y}_i\right) \tag{1.3}$$

minimize etmektir. Bu kriter *en küçük kareler yöntemi* olarak adlandırılır. Bu yöntemin avantajları aşağıda verilmiştir.

a) En küçük kareler yönteminin cebirsel işlemleri oldukça basittir.

b) Hataların karesinin alınması işaret problemini ortadan kaldırır.

c) Kare alma işlemi büyük hata terimlerini daha da büyüterek vurgulanmasını sağlar.

Bu kriterin uygulanması için çalışmalar yapılırken eğer mümkünse büyük değerli hatalar modelde yapılacak düzeltmelerle ortadan kaldırılır. Bu nedenle kriter uygulanırken tüm noktalar dikkate alınmaktadır. Şekil 1.3' de verilen uyumlar bu kritere göre değerlendirildiğinde (*a*)' daki uyum (*b*)' ye göre tercih edilir.

1.2 EN KÜÇÜK KARELER TAHMİNLEME YÖNTEMİ

Gözlemlere en iyi uyumu sağlayan, doğrusal ya da doğrusal olmayan modelin uyarlanması için bir çok yöntem vardır. En küçük kareler (EKK) yöntemi bunlar arasında en yaygın olarak kullanılanıdır. Burada açıklanması gereken önemli bir nokta vardır. EKK yöntemi gözlemlere en iyi uyum sağlayacak matematiksel modeli bulmaz, matematiksel modeli belirlenmiş bir durum için, verilere en iyi uyum sağlayacak parametre tahminlerini yapar. Örneğin, gerçek modelin üçüncü dereceden bir polinom olduğu durumda kullanılacak modelin üçüncü dereceden bir polinom olması gerektiğini belirtmez. Eğer bu durumda uyumu sağlanacak bir doğru söz konusu ise EKK yöntemi gözlemler ile doğrunun en iyi uyumunu sağlayacak parametreleri tahminler.

EKK yöntemi, belirlenen modelde hata kareler toplamını minimum yapan ve parametrelerin sapmasız minimum varyanslı tahminlerini elde eden yöntemdir.