# REGRESYON ANALİZİ
# BÖLÜM 5-6

Yayın Tarihi: 03-11-2007
Revizyon No:0

1

# 5. E.K.K. REGRESYONUNDA KARŞILAŞILAN PROBLEMLER VE BAZI KONU BAŞLIKLARI

2

- EN KÜÇÜK KARELERDE KARŞILAŞILAN PROBLEMLER
  - EKK da karşılaşılan problemler:
    - normallik,
    - sabit varyans
    - hataların bağımsızlığı
    - etkili gözlemler,sapan gözlemler,
    - modelin fonksiyonel biçiminin zayıf belirlenmesi,
    - bağımsız değişkenler arasındaki çoklu doğrusal bağlantı
    - bağımsız değişkenlerdeki hatalardır.

3

- ROBUST REGRESYON
  - Varsayımlar gerçekleşmediğinde en küçük kareler regresyonuna alternatif olarak "*robust regresyon*" yöntemleri kullanılabilir.
  - Robust regresyon,parametrik modelin varsayımlarının gerçekleşmemesi durumunda tahminlerin duyarlılığını azaltmak için tasarlanmış istatiksel prosedürlerin genel bir sınıfıdır.
  - Bir robust regresyon prosedürü,büyük artıkların ağırlıklarının azaltılması ile, bu tip hataların etkisini azaltır.
  - Bu artıkların etkisi, artık kareler toplam yerine mutlak artıkların toplamının minimize edilmesi ile azaltılabilir.
  - Genel anlamda, sapan ve etkili gözlemlerin tespit edilmesi için kullanılan prosedürler, robust regresyonun bir parçası olarak ele alınabilir.

4

- NORMALLİK VARSAYIMI

  - Hataların normal dağılış göstermesi, regresyon parametrelerinin tahminlenmesi ve toplam varyasyonun parçalanması için gerekli değildir.
  - Normallik varsayımı, yalnızca parametrelerin güven aralıklarının kurulmasında ve anlamlılık testlerinde gerekmektedir.

5

- NORMALLİK VARSAYIMININ KONTROLÜ

  - Normallik varsayımının kontrolünde,
    - artıkların plotları,
    - çarpıklık katsayısı ve
    - basıklık katsayısı,

    yardımcı olurlar.
  - Örnek hacmi yeterince büyük olduğunda, artıkların
    - frekans dağılışı,

    simetri ve basıklık için karar vermede kullanılabilir.
  - Normalliğin sağlanması durumunda bir doğru veren,
    - bir tam normal veya yarı normal plotun

    kullanımı ise daha kolaydır.

6

- TAM VE YARI NORMAL PLOT
  - Bu plotlar, verilerden elde edilen, sıralanmış artıklar ile standart normal dağılıştan elde edilen sıralanmış gözlemlerin beklenen değerlerini karşılaştırır.
  - Tam-normal plot, artıkları olduğu gibi, yarı-normal plot ise artıkların mutlak değerlerini kullanır.

7

- NORMALLİK VARSAYIMININ SAĞLANMASINDA KULLANILAN YÖNTEMLER
  - Bağımlı değişkenin başka bir forma transformasyonu, normallik varsayımının sağlanması için sıkça uygulanan bir yöntemdir.
  - İstatistik teorisine göre, orjinal bağımlı değişkenin dağılışı bilindiğinde bu tip bir transformasyon yapılabilir.
  - Örnek verileri, normalliğin sağlanması için uygun olan transformasyonun bulunması ile gerekli bilgiyi sağlar.
  - Artık plotları, uygun transformasyon ile ilgili fikir verebilir.
  - Ayrıca birden çok sayıda transformasyon denerek normallik varsayımını en iyi sağlayan bir tanesi seçilebilir.

8

- VARYANS HETEROJENLİĞİ
  - Sabit varyans varsayımı, bağımlı değişkendeki her gözlemin, eşit miktarda bilgiyi içerdiği anlamına gelmektedir.
  - Diğer bir ifadeyle, sıradan en küçük karelerdeki (SEKK) tüm gözlemler eşit ağırlıktadır.
  - Diğer taraftan, varyans heterojenliğinin varlığı durumunda, bazı gözlemler diğerlerinden daha fazla bilgiyi içermektedir.
  - SEKK tahminleyicilerinin minimum varyans özelliği bu varsayım ile doğrudan bağımlıdır.

9

- VARYANS HETEROJENLİĞİ
  - Gözlemler SEKK ile eşit ağırlıklandırıldığında, eğer varyanslar eşit değilse, parametre tahminleri minimum varyans özelliğini taşımazlar.
  - SEKK de heterojen varyanslar, doğrudan tahminlerin hassasiyet kaybına neden olmaktadır.
  - Bu hassasiyet kaybı ancak, heterojen varyanslar dikkate alınarak elde edilen tahminlerin hassasiyetleri ile dikkate alınmadan elde edilen tahminlerin hassasiyetleri karşılaştırılarak görülebilir.
  - Normal olmama durumunda olduğu gibi, heterojen varyansın da verilerden kaynaklanması beklenir.

10

- VARYANS HETEROJENLİĞİNİN BELİRLENMESİ
  - Normal olmayan dağılışları veren durumlar, varyans heterojenliğine de neden olurlar.
  - Bunun nedeni, normal olmayan diğer dağılışlarda varyansın, dağılış ortalaması ile ilişkili olmasıdır.
  - Bununla beraber, verilerin kendi içindeki gruplar ayrı ayrı normal dağılış gösterseler bile, bu dağılışların varyansları gruptan gruba değişiklik gösterebilir.
  - Genel olarak, büyük varyanslar, büyük ortalamalara sahip gruplar ile birlikte ortaya çıkmaktadır.
  - Çeşitli artık plotları, varyans heterojenliğini tespit etmek için kullanılabilir.

11

- FARKLI VARYANSLILIĞIN ÖNLENMESİ
  - Varyans heterojenliğinin önlenmesi için iki ayrı yaklaşım vardır.
    - Bağımlı değişkenin transformasyonu ve
    - Ağırlıklı en küçük kareler yöntemleridir.
  - Bağımlı değişkenin transformasyonu yaklaşımı daha yaygındır.
  - Buradaki transformasyon, transforme edilmiş ölçekteki varyansı homojen yapacak şekilde seçilir.

12

- FARKLI VARYANSLILIK VARSAYIMININ ÖNLENMESİ
  - Bağımlı değişkenin olasılık dağılışı hakkındaki bir ön bilgi veya ortalama ile varyans arasındaki ilişki hakkındaki amprik bilgi, transformasyonunun şekli hakkında bilgi verir.
  - Ağırlıklı en küçük kareler ise, bağımlı değişkenin orijinal metriğini kullanır. Ancak, içerdiği bilginin miktarına göre her bir gözlemi ağırlıklandırır.

13

- HATALAR ARASINDAKİ KORELASYON (OTOKORELASYON)
  - Artıkların arasındaki korelasyonun (otokorelasyon) birçok kaynağı olabilir.
  - Bir zaman serisinde toplanan verilerin korelasyonlu hatalara sahip olması sıkça rastlanan bir durumdur.
  - Bir zaman noktasındaki gözlem ile ortaya çıkan hata, bir önceki gözlem ile ortaya çıkan hata ile korelasyonlu olma eğilimi gösterebilir.
  - Korelasyonlu hataların SEKK sonuçları üzerindeki etkisi, varyans heterojenliğinin etkisi gibi, tahminlerin hassasiyetinde kayba neden olmasıdır.

14

- OTOKORELASYONUN NEDENLERİ
  - Verilerin yapısı, bazen korelasyonlu hataların varlığının önerebilir.
    - Zaman sırasında toplanmış herhangi bir veri seti, korelasyonunun önemsiz olduğu gösterilmedikçe zaman serisi gibi görülmelidir.
    - Bir deneyin randomizasyonundaki yetersizlikten veya randomizasyon planındaki bir hatadan kaynaklanan korelasyonlu hataların tespit edilmesi ise zordur.

15

- OTOKORELASYONUN OLUŞTURDUĞU PROBLEMLER
  - Eğer regresyon modelindeki hatalar pozitif otokorelasyona sahip ise,
    - EKK parametre tahminleri sapmasızdır fakat minimum varyans özelliğini kaybetmiştir.
    - MSE değeri gerçek hata varyansını olduğundan küçük olarak tahminlemektedir.
    - EKK ile tahminlenen s($b_i$) değerleri gerçek değerinden oldukça küçüktür.
    - Güven aralıkları ve *t* ile *F* dağılımlarını kullanan testler artık kullanışlı değildir.

16

- OTOKORELASYON: TANI VE ÖNLEM

- Deneysel çalışmalarda, aşırı derece küçük artık varyansları bir fikir verebilir.
- Zaman sırası verilerinde ise, verilerin toplandıkları sıraya bağlı olarak, artıkların plot edilmesi otokorelasyonunun olup olmadığını belirlemek amacıyla kullanılabilir.
- Genelleştirilmiş en küçük kareler, korelasyonlu hatalara sahip verilerin analizi için genel bir yaklaşımdır.
- Bu yöntemde ağırlıklı en küçük karelerde (AEKK) olduğu gibi artıkların başlangıç varyans-kovaryans matrisi kullanmaktadır.
- Ancak artıklar arasındaki kovaryanslar genellikle bilinmemektedir ve verilerden tahminlenmesi gerekmektedir.
- Eğer korelasyon yapısı basit değilse tahminlenmesi zordur ve korelasyon matrisinin zayıf tahminlenmesi, hassasiyet kaybına neden olur.

17

- ETKİLİ VERİ NOKTALARI VE SAPANLAR

- SEKK yöntemi, her bir gözleme eşit ağırlık vermektedir.
- Fakat, her gözlem, çeşitli en küçük kareler sonuçları üzerinde eşit etkiye sahip değildir.
- Örneğin, basit doğrusal regresyon probleminde eğim en çok, ortalamadan en uzakta bulunan bağımsız değişken değerlerine sahip gözlemlerden etkilenmektedir.
- Diğer veri noktalarından uzak olan tek bir nokta, regresyon sonuçları üzerinde, hemen hemen tüm diğer noktaların toplamı kadar bir etkiye sahip olabilir.

18

- TANIMLAR
  - Bu tip gözlemler *etkili gözlemler* olarak adlandırılır.
  - *Sapan terimi*, veri setindeki kalan gözlemlerle karşılaştırıldığında tutarsız olan gözlem için kullanılır.
  - Bir gözlem, bağımlı değişken veya bir ya da daha fazla sayıdaki bağımsız değişkenin, beklenen limitleri dışında değerler alması nedeniyle sapan gözlem olabilir.
  - *Potansiyel etkili gözlem* terimi, bir ya da daha fazla sayıdaki bağımsız değişkendeki sapan gözlem için kullanılır.

19

- SAPAN GÖZLEM
  - Sapan terimi, bağımlı değişken değerleri içerisinde, diğer gözlemler ile karşılaştırıldığında tutarsız olan gözlem için kullanılır.
  - Artıklardaki sapan terimi ise, gözlenen artığın, beklenen varyasyonundan daha büyük olduğu veri noktası için kullanılır.
  - Burada ″Sapan″ terimi bağımlı değişken veya artığın değerine karşılık gelecek şekilde kullanılmıştır.

20

- SAPAN GÖZLEM
  - Bir veri noktası, çalışmanın yürütülmesindeki hatalardan veya veri noktasının başka bir populasyondan gelmesinden dolayı sapan ya da potansiyel etkili nokta olabilir.
  - Kullanılan model süreci yeterli derecede temsil etmiyorsa, aslında doğru olan bir nokta, sapan artığa sahip bir sapan olarak görülebilir.
  - Diğer taraftan, gerçek bir sapan, potansiyel etkili nokta ise sapan artığa sahip olmayabilir.

21

- ETKİLİ GÖZLEM
  - Etkili veri noktaları regresyon doğrusunu zorlama eğilimindedirler ve bu yüzden küçük artılara sahiptirler.
  - Etkili noktaların ve sapanların tespit edilmesi gerekmektedir.
  - Yalnızca birkaç gözlemin etkisinde olan regresyon sonuçları güvenilir olmazlar.
  - Doğrulanması gereken ilk şüphe, bu veri noktalarının doğru olup olmadığıdır.
  - Açık bir şekilde farkedilebilen hatalar, eğer mümkünse düzeltilmelidir.
  - Aksi halde, veri setinden çıkarılmalıdır.

22

- ETKİLİ GÖZLEM
  - Açık bir şekilde hata olarak tanımlanamayan veya üzerinde çalışılan sistem hakkında içerdiği bilgi nedeniyle, dikkatli bir şekilde incelenmesi gereken doğru bir nokta olabilir.
  - Bu noktalar üzerinde çalışılan model veya tasarımdaki yetersizlikleri mi yansıtmaktadır?
  - Sapanlar ve etkili veri noktaları, rastgele çıkartılmamalıdır.
  - Bir sapan, çalışmadaki en çok bilgi verici gözlem bile olabilir.

23

- POTANSİYEL ETKİLİ GÖZLEM: TEŞHİS YÖNTEMLERİ
  - Potansiyel olarak daha etkili olan noktalar, izdüşüm matrisi **H**'nin köşegen elemanlarının kontrol edilmesiyle bulunabilir.
  - **H** matrisinin köşegen elemanları, örnek noktaları ile *X*-uzayının merkezi arasındaki Öklid Uzaklığının ölçüsünü verir.
  - Ancak bir potansiyel etkili noktanın, gerçekte etkili olup olmadığı, her bir veri noktasının regresyon sonuçları üzerindeki etkisi doğrudan ölçülerek bulunur.

24

- SAPANLAR:
  TEŞHİS
  YÖNTEMLERİ
    - Sapanlar, gözlenen artıkların analizi ile tespit edilebilir.
    - Genellikle artıkların genel bir varyansa sahip olmaları için önce standardize edilmeleri tavsiye edilir.
    - Birkaç standart sapma büyüklüğünde olan bir artık, dikkatlice gözden geçirilmesi gereken bir veri noktasına işaret etmektedir.
    - Normallik ve varyans homojenliği varsayımlarını sağlayıp sağlamadığını tespit etme amacı ile kullanılan artıklarının plotları, sapanları tespit etmede de etkilidirler.

25

- MODEL
  YETERSİZLİKLERİ
    - Eğer model doğru değilse ise, SEEK tahminleyicileri sapmalı olurlar.
    - Önemli bir bağımsız değişken modelde ihmal edilmişse, artık kareler ortalaması, $\sigma^2$'nin (pozitif) sapmalı bir tahmini olur.
    - Ayrıca, ihmal edilen değişken, modeldeki diğer değişkenlere ortogonal değilse regresyon katsayıları da sapmalı olur.
    - Bağımsız değişkenlerin yalnızca birinci kuvvetlerini kullanan genel doğrusal model, *Y*'nin her bir bağımsız değişken ile olan ilişkinin doğrusal olduğunu ve her bir bağımsız değişkenin etkisinin diğer değişkenlerden bağımsız olduğunu varsayar.

26

- MODEL YETERSİZLİKLERİ
  - Önemli olan daha yüksek-dereceli polinomiyal terimlerin (etkileşim terimleri de dahil olmak üzere) ihmal edilmesinin etkisi, önemli bir bağımsız değişkenin ihmal edilmesinin etkisi ile aynıdır.
  - Gerçek ilişkinin doğrusal olmadığı durumlarda doğrusal bir model kullanılmasındaki temel düşünce, doğrusal olmayan fonksiyona, istenilen bir doğruluk derecesinde, uygun sayıda polinomial terimleri olan doğrusal bir model ile yaklaşılabileceğidir.
  - Böylece, doğrusal model, ilgilenilen sınırlı bir bölgede tatmin edici bir yaklaşım sağlar.
  - Yaklaşımın yeterli olmaması durumunda, en küçük kareler sonuçları, önemli bir değişkenin ihmal edilmesi durumunda olduğu gibi sapmalı olacaktır.

27

- MODEL YETERSİZLİKLERİ
  - Süreç modellerine doğrusal modeller ile yaklaşılması *Cevap Yüzeyi Yöntem Bilimi* ile ilgilidir.
  - Model yetersizliklerinin tespit edilmesi,
    - problemin yapısına ve
    - sistemden elde edilebilen bilginin miktarına bağlıdır.
  - Artık kareler ortalamasındaki sapma ve dolayısıyla ihmal edilen terimin varlığı, $\sigma^2$'nin bağımsız bir tahmini elde edilebiliyorsa tespit edilebilir.
  - Dikkate alınmayan yüksek-dereceli polinomiyal terimler uygun artık plotları ile kolayca görülebilir.
  - Tümüyle ihmal edilmiş bağımsız değişkenlerin tespit edilmesi ise daha zordur.

28

- MODEL YETERSİZLİKLERİ
  - Doğrusal yaklaşımlara alternatif olarak daha gerçekçi olan doğrusal olmayan modeller formüle edilebilir.
  - Bazı doğrusal olmayan modeller, bağımlı değişkenin uygun bir transformasyonu ile doğrusallaştırılabilir.
  - Bu tip modeller ″aslında doğrusal modeller″ olarak adlandırılır.
  - Transformasyondan sonra hatalar üzerindeki varsayımlar sağlanıyorsa, doğrusallaştırılmış modele SEKK uygulanabilir.

29

- DOĞRUSAL BAĞLANTI PROBLEMİ
  - **X** matrisinin tekilliği, **X** in bazı sütunlarının bir doğrusal fonksiyonunun kesin olarak sıfır vektörüne eşit olması durumunda çıkar.
  - Tekillik EKK analizinde $(X^TX)^{-1}$ var olmadığında belli olur.
  - Daha problemli bir durum ise **X** matrisi tekile yakın olduğunda, yani vektörlerin doğrusal fonksiyonu sıfıra yakın olduğunda ortaya çıkar.
  - Değişken sayısının gereğinden fazla olması (aynı bilgiyi değişik formlarda açıklamaktadır), **X** in tekile yakın olmasına neden olur.

30

- YAKLAŞIK-TEKİLLİK DURUMU
  - Yaklaşık-tekillik durumlarında normal eşitliklerin tek bir çözümcü mevcuttur.
  - Ancak bu çözüm oldukça kararsızdır.
  - Böyle bir durumda **y** veya **X** deki küçük değişikler (şansa bağlı gürültü), regresyon katsayılarının tahminlerinde büyük değişmelere neden olabilir.
  - Ayrıca regresyon katsayılarının varyansları oldukça büyük olurlar.
  - **X** in yaklaşık-tekil olmasından kaynaklanan problemlere yaklaşık doğrusal bağlantı problemi olarak adlandırılır.

31

- ÇOKLU DOĞRUSAL BAĞLANTININ BELİRTİLERİ VE TESBİT EDİLMESİ
  - Doğrusal bağlantının varlığını gösteren ipuçları:
    - regresyon katsayıları için mantıksız değerler,
    - büyük standart hatalar,
    - model mantıklı bir uyum sağladığı halde anlamsız olan kısmi regresyon katsayıları ve
    - regresyon sonuçlarında önemsiz görülen ancak önemli olduğu bilinen değişkenler.
  - Doğrusal bağlantının tespit edilmesi doğrudan,
    - **X** in tekil değer ayrışımı veya
    - $\mathbf{X}^T\mathbf{X}$ in özdeğer-özvektör analizi ile sağlanabilir.

32

- ÇOKLU DOĞRUSAL BAĞLANTI PROBLEMİNİN ÇÖZÜMÜ
  - Doğrusal bağlantı probleminin çözümü, modelin kurulmasındaki amaca bağlıdır.
  - Eğer amaç kestirim ise, örnek *X*-uzayında doğrusal bağlantı ciddi bir problem yaratmaz.
  - Eğer amaç regresyon katsayılarının tahminlenmesi ise, parametrelerinin tahminleri, parametre değerinden büyük farklılıklar gösterebilir; hatta yanlış işarete bile sahip olabilir.

33

- ÇOKLU DOĞRUSAL BAĞLANTI PROBLEMİNİN ÇÖZÜMÜ
  - Sapmalı regresyon yöntemlerinden birisi kullanılabilir.
    - Daha iyi bir çözüm ise, eğer mümkünse, yeni veya ek verilerin elde edilmesi ve böylece örnek **X**-uzayının yaklaşık-tekilliği ortadan kaldırmak için genişletilmesidir.
  - Söz konusu amaç bir sistemdeki önemli değişkenleri tanımak veya sistemi modellemek ise, kuvvetli doğrusal bağlantının olması durumunda regresyon sonuçları fazla yardımcı olmazlar ve yanıltıcı olabilirler.
    - Burada, değişkenlerin korelasyon yapısını anlamak ve bağımlı değişkenin bu yapıya nasıl uyumunun yapılacağını araştırmak daha verimli olur.
    - Ana bileşenler analizi ve ana bileşenler regresyonu bu yapıyı anlamada yardımcı olurlar.

34

- BAĞIMSIZ DEĞİŞKENLERDEKİ ÖLÇÜM HATALARI

  - Ölçüm hassasiyetine bağlı olarak, ısı basınç, kişinin yaşı gibi bağımsız değişkenler ölçülerken hata yapılabilir.

    **(5.1)** $\delta_i = X_i^* - X_i$
  - Basit regresyon modeli için,

    **(5.2)** $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
  - Sadece $X_i^*$ gözlendiği için

    **(5.3)** $Y_i = \beta_0 + \beta_1 (X_i^* - \delta_i) + \varepsilon_i$

    **(5.4)** $Y_i = \beta_0 + \beta_1 X_i^* + (\varepsilon_i - \beta_1 \delta_i)$

35

- BAĞIMSIZ DEĞİŞKENLERDEKİ ÖLÇÜM HATALARI

  - Bu model bağımsız değişkeni $X_i^*$ ve hata terimi $\varepsilon_i - \beta_1 \delta_i$ olan bir regresyon modeli olarak görülse de değildir.
  - Bağımsız değişken şans değişkeni olup hata terimi ile ilişkilidir.
  - $X$ in rassal değişken olduğu modeller için de rassal bağımsız değişken hata teriminden bağımsız olmalıdır.

36

- BAĞIMSIZ DEĞİŞKENLERDEKİ ÖLÇÜM HATALARI
  - Modeli belirlemek için aşağıdaki varsayımlar yapılabilir.

    **(5.5a)** $E(\delta_i) = 0$

    **(5.5b)** $E(\varepsilon_i) = 0$

    **(5.5c)** $E(\delta_i \varepsilon_i) = 0$
  - Bunun sonucu olarak,

    $$E(X_i^*) = E(X_i - \delta_i) = X_i$$
  - Kovaryans ise,

    **(5.6)** $\sigma(X_i^*; \varepsilon_i - \beta_1 \delta_i) = -\beta_1 \sigma^2(\delta_i)$

    ☞ İspat 40
  - Eğer X ve Y arasında bir regresyon ilişkisi varsa bu kovaryans sıfır olamaz.

37

- ÖLÇÜM HATALARININ SONUÇLARI VE ÖNLEMLER
  - Eğer model (5.4) e EKK uygulanırsa $b_i$ parametre tahminleri sapmalı olacaktır ve kararlılık özelliğini kaybedeceklerdir.
  - Sapmasız tahminleyicileri elde etmek için yaklaşımlar:
    - $\delta_i$ nın dağılımı ve $\delta_i$ ile εi arasındaki kovaryans üzerine güçlü varsayımlar yapılır.
    - $X$ in gerçek değeri ile ilişkisi bilinen fakat ölçüm hatasına sahip olmayan değişkenleri ( alet – instrumental) kullanmak.
  - Alet değişkenlerin kullanılması regresyon parametrelerinin tutarlı tahminlerinin elde edilmesine imkan sağlar.

38

- X İN ÖLÇÜM HATASINA SAHİP OLDUĞU DURUM İLE X İN ŞANS DEĞİŞKENİ OLDUĞU DURUM ARASINDAKİ FARK:
  - X şans değişkeni ise analizcinin kontrolü dışındadır ve deneyden deneye şansa bağlı olarak değişir.
  - Bu durumda X ölçüm hatasına sahip değilse kesin olarak (verilen bir deneme için) ölçümlenir.
  - Bazı araştırmalarda ise X belirli değerlerde sabitlenir. Örneğin ısı ve basınç gibi. Fakat bu değişkenleri istenen hassasiyette sabitlemek mümkün olmadığından X ölçüm hatalı olarak adlandırılır.

39

- PARAMETRELERİN ORTAK GÜVEN ARALIKLARI
  - Parametre vektörünün dağılımının,
    $$\mathbf{b} \sim N\left(\boldsymbol{\beta}, \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\sigma^2\right)$$
    varsayımı altında
  - Tüm parametreler için %100(1-$\alpha$) güven seviyesinde bir ortak güven aralığı,

$$\mathbf{(5.7)}\quad (\boldsymbol{\beta}-\mathbf{b})^T\mathbf{X}^T\mathbf{X}(\boldsymbol{\beta}-\mathbf{b}) \le ps^2F(p, n-p, 1-\alpha)$$

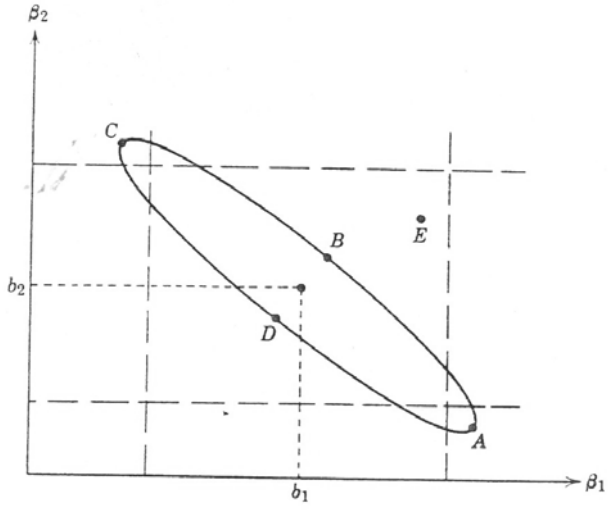
  - Denklem *p* boyutlu uzayda eliptik bir eğrinin konturunu tanımlar.
  - Bu yaklaşım sadece *p* değerinin küçük (2,3,4) olduğu durumlar için faydalıdır.

40

- PARAMETRELERİN ORTAK GÜVEN ARALIKLARI
  - Şekil 5.1 iki parametreli mümkün bir durum örneğidir.
  - Parametreler $\beta_1$ ve $\beta_2$ için %95 güven bölgesi ince uzun elips ile gösterilmiştir.
  - Elips içindeki ($\beta_1$, $\beta_2$) değerleri parametrelerin ortak olarak alabilecekleri değerlerdir.
  - Bu yaklaşım $b_1$ ve $b_2$ tahminleri arasındaki korelasyonu dikkate alır.

41

- Şekil 4.1




42

- PARAMETRELERİN ORTAK GÜVEN ARALIKLARI

  - Parametreler $\beta_1$ ve $\beta_2$ için bireysel ayrı ayrı %95 güven aralıkları kesikli kare ile belirtilmiş olup eş anlı yorumlanması yanlıştır.
  - Örneğin $E$ noktası $(\beta_1, \beta_2)$ uygun değerlere sahiptir.
  - Bununla birlikte ortak güven bölgesi böyle bir noktanın uygun olmadığını belirtmektedir.
  - Parametre sayısı ikiden fazla olduğunda yorumlamak zordur.

43

- PARAMETRELERİN ORTAK GÜVEN ARALIKLARI

  - Böyle durumlarda eliptik bölgenin ana eksenlerinin uç noktalarının koordinatlarını bulmak bir çözüm olabilir.
  - Başarmak için güven konturu bulunmalı ve kanonik yapıya indirgenmelidir.
  - Bu amaçla kullanılabilecek yöntemler; ana bileşenler regresyonu ve latent kök regresyonudur.
  - Eğer model,
    $E(Y - \bar{Y}) = \beta_1(X_1 - \bar{X}_1) + \cdots + \beta_k(X_k - \bar{X}_k)$
    şeklinde yazılırsa $\beta_0$ parametresini içermeyen ortak güven bölgesi elde edilir.
  -

44

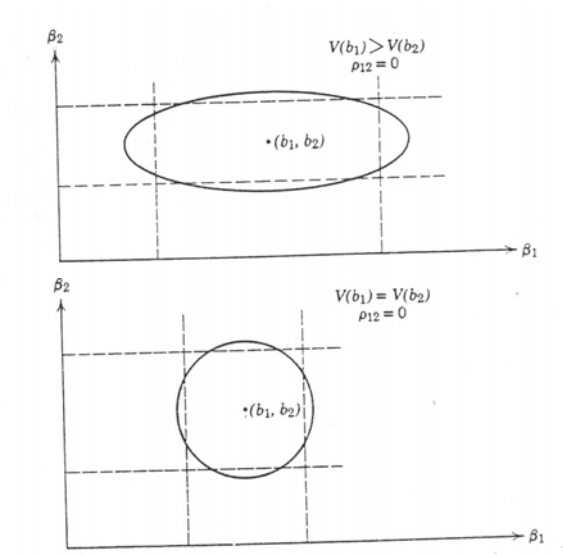
- ORTAK GÜVEN BÖLGELERİ İÇİN DİKKAT EDİLMESİ GEREKEN KONULAR
  - Parametre tahminlerinin,
    - o Varyansları $V(b_i)$
    - o Korelasyonları $\rho_{ij}$

    **(5.8)** $$\rho_{ij} = \frac{Cov(b_i, b_j)}{[V(b_i)V(b_j)]^{1/2}}$$

    incelenmelidir.
  - Mümkün durumlara örnekler:
    - o $V(b_i) > V(b_j)$ ve $\rho_{ij} \neq 0$ Şekil 5.1
    - o $V(b_i) > V(b_j)$ ve $\rho_{ij} = 0$ Şekil 5.2a
    - o $V(b_i) = V(b_j)$ ve $\rho_{ij} = 0$ Şekil 5.2b

45

- Şekil 5.2




46

- ORTAK GÜVEN BÖLGELERİ İÇİN DİKKAT EDİLMESİ GEREKEN KONULAR
  - Parametreler $\beta_0$ ve $\beta_1$ için bireysel ayrı ayrı %95 güven aralıkları birlikte yorumlandığında eğer birbirinden bağımsız iseler $(0.95)^2=0.9025$ güven olasılığına sahiptirler.
  - Bununla birlikte aynı veri seti kullanılarak tahminlendiğinden birbirinden bağımsız değildirler.
  - Yorumların doğru olasılığını belirlemek amacıyla kullanılabilecek bir yöntem *Bonferroni ortak güven aralıkları* yöntemidir.

47

- BONFERRONİ ORTAK GÜVEN ARALIKLARI
  - Verilerin analizi genellikle bir dizi tahmin ya da testin gerçekleştirilmesini gerektirir.
  - Gerçekleştirilen test ya da tahminlerin bütün kümesine ait doğrulukla ilgili bilgiye gereksinim vardır.
  - İlgilenilen tahminler (ya da testler) kümesi, *tahminler* (ya da test) *ailesi* olarak adlandırılır.

48

- BONFERRONİ ORTAK GÜVEN ARALIKLARI
  - Bonferroni prosedüründe, aile güven katsayısını en az 1-α olacak şekilde düzenlemek amacıyla bireysel güven katsayıları 1-α den büyük olacak şekilde düzeltilir.
  - İki parametreli (iki eşanlı güven aralığı) için Bonferroni eşitsizliği:

**(5.9a)** $P(\overline{A}_1 \cap \overline{A}_2) \geq 1 - 2\alpha$

  - *k* parametreli (*k* adet eşanlı güven aralığı) için Bonferroni eşitsizliği:

**(5.9b)** $P\left(\bigcap_{i=1}^{k} \overline{A}_i\right) \geq 1 - k\alpha$

  - ☞ **İspat 39**

49

- YUVARLAMA HATALARI
  - Bir regresyon modelinde yalnızca bir veya iki bağımsız değişken olduğunda parametre tahminlerinin doğrudan hesaplanması genellikle hesaplamalarda fazla sorun çıkarmamaktadır.
  - İkiden fazla değişken ve fazla sayıda verilerin bulunduğu problemlerde ise elde edilen sonuçlar yuvarlama hataları nedeniyle tamamen geçersiz olabilmektedir.

50

- YUVARLAMA HATALARININ NEDENİ
  - İki ana nedeni:
    - Regresyon hesaplamalarına dahil edilen sayılar birbirinden çok büyük farklılıklar gösterebilir. Örneğin 52.793, -943 ve 6 sayılarının aynı hesaplamaya dahil edilmesi gibi
    - Tersi alınacak matris yaklaşık tekil matris olabilir. Matrisin determinantı, hesaplamalara dahil edilen diğer sayılarla kıyaslandığında oldukça küçük ise yuvarlama problemi büyük bir olasılıkla ortaya çıkacaktır. $|\mathbf{X}^T\mathbf{X}|$ çok küçük bir değer olduğunda $\mathbf{X}^T\mathbf{X}$ matrisine kötü şartlanmış (ill conitioned) matris denir.

51

- YUVARLAMA HATALARININ ETKİSİNİ AZALTMAK İÇİN NE YAPILIR?
  - Regresyon değişkenlerinin *standartlaştırılmasıyla* regresyon problemi korelasyonları içeren bir forma dönüşür ve böylece hesaplamalara dahil edilen bütün sayılar -1 ile 1 arasında değişir.
  - Sayıların standart bir forma dönüşmesiyle yuvarlama hatalarının olumsuz etkileri de en aza indirilmiş olur.

52

- MERKEZLENMİŞ DEĞİŞKENLERE GÖRE MODEL
  - Genel doğrusal model

  **(5.10)** $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon$
  - Bu model bağımsız değişkenlerin ortalamadan farklarına göre yazılarak *merkezlenmiş* değişkenlere göre elde edilir.

  **(5.11)** $Y_i = \{\beta_0 + \beta_1 \bar{X}_1 + \cdots + \beta_k \bar{X}_k\} + \beta_1 (X_{i1} - \bar{X}_1) + \cdots + \beta_k (X_{ik} - \bar{X}_k) + \varepsilon$
  - Burada $x_j = X_{ij} - \bar{X}_j$ merkezlenmiş değişkendir ve $\beta_0' = \beta_0 + \beta_1 \bar{X}_1 + \cdots + \beta_k \bar{X}_k$ alınarak model,

  **(5.12)** $Y_i = \beta_0' + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$

53

- MERKEZLENMİŞ DEĞİŞKENLERE GÖRE MODEL
  - Merkezlenmiş değişkenlerin ortalaması sıfırdır.

    $\bar{x}_j = 0$
  - $\beta_0'$ parametresinin tahminleyicisi daima bağımlı değişkenin aritmetik ortalamasına eşittir.

    **(I41.1)** $b_0' = \bar{Y}$

    ☞ İspat 41
  - Sonuç olarak (5.12) modeli

    **(5.13)** $Y_i - \bar{Y} = \beta_1 (X_{i1} - \bar{X}_1) + \cdots + \beta_k (X_{ik} - \bar{X}_k) + \varepsilon$

54

- MERKEZLENMİŞ MODEL İÇİN İÇ ÇARPIM MATRİSİ

  - Bu modelde $k$=2 durumu için için $\mathbf{X}^T\mathbf{X}$ matrisi,

    **(5.14)** $$\mathbf{X}^T\mathbf{X} = \mathbf{S} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$$

  - Burada,

    **(5.15)** $$S_{lj} = \sum_{i=1}^{n} \left(X_{il} - \overline{X}_l\right)\left(X_{ij} - \overline{X}_j\right)$$

55

- STANDARTLAŞTIRILMIŞ DEĞİŞKENLERE GÖRE MODEL

  - Merkezlenmiş veriler ölçeklenerek standartlaştırılmış değişkenler elde edilir.

    **(5.16a)** $$z_{ij} = \frac{\left(X_{ij} - \overline{X}_j\right)}{S_{jj}^{1/2}} = \frac{x_{ij}}{S_{jj}^{1/2}} \;,\; j = 1,\ldots,k$$

    **(5.16b)** $$y_i = \frac{\left(Y_i - \overline{Y}\right)}{S_{yy}^{1/2}} \qquad i = 1,\ldots,n$$

  - Burada, $S_{yy} = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$
  - Standartlaştırılmış değişkene göre model,

    **(5.17a)** $$y_i S_{yy}^{1/2} = \beta_1 S_{11}^{1/2} z_{i1} + \cdots + \beta_k S_{kk}^{1/2} z_{ik} + \varepsilon$$

    **(5.17b)** $$y_i = \alpha_1 z_{i1} + \cdots + \alpha_k z_{ik} + \varepsilon$$

56

- STANDART MODEL İÇİN İÇ ÇARPIM MATRİSİ

- Standart modelde $k$=2 durumu için için $\mathbf{X}^T\mathbf{X}$ matrisi,

**(5.18)** $$\mathbf{X}^T\mathbf{X} = \mathbf{R} = \begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix}$$

- Burada,

**(5.19a)** $$r_{12} = \frac{S_{12}}{(S_{11}S_{22})^{1/2}} = r_{21}$$

☞ İspat 42

- Bağımlı değişkenle bağımsız değişken arasındaki korelasyon,

**(5.19b)** $$r_{jy} = \frac{S_{jy}}{(S_{jj}S_{yy})^{1/2}}$$

- Burada,

**(5.20)** $$S_{jy} = \sum_{i=1}^{n}(X_{ij} - \bar{X}_j)(Y_i - \bar{Y})$$

57

- STANDARTLAŞTIRILMIŞ DEĞİŞKENLERE GÖRE MODEL

- Burada yeni modelin parametreleri ve tahminleyicileri,

**(5.21a)** $$\alpha_j = \frac{\beta_j S_{jj}^{1/2}}{S_{yy}^{1/2}}$$

**(5.21b)** $$a_j = \frac{b_j S_{jj}^{1/2}}{S_{yy}^{1/2}}$$

- Orijinal parametre cinsinden,

**(5.22a)** $$\beta_j = \frac{\alpha_j S_{yy}^{1/2}}{S_{jj}^{1/2}}$$

**(5.22b)** $$b_j = \frac{a_j S_{yy}^{1/2}}{S_{jj}^{1/2}}$$

☞ İspat 43

58

# 6. ARTIKLARIN ANALİZİ VE REGRESYON TANILARI

59

- REGRESYON TANILARI NEDİR?
  - Regresyon tanıları, genel olarak regresyon analizinde ortaya çıkan problemleri belirlemek için kullanılan teknikleri ifade eder.
  - Problemlerin nedeni,
    - Model ya da
    - Veri seti

    kaynaklı olabilir.

60

- REGRESYON TANI YÖNTEMLERİ
  - İncelenecek regresyon tanıları,
    - Grafiksel yöntemler:
      - varsayımlardaki tutarsızlıkları,
      - sapan gözlemleri
      - model yetersizliklerini

      belirlemek için.
    - Tanı istatistikleri:
      - Sapan gözlemleri
      - Yüksek etki noktalarını
      - Etkili gözlemleri
      - Çoklu doğrusal bağlantıyı

      belirlemek için.

61

- TEMEL KAVRAMLAR

  - *Sapan* (outliers) *gözlem*,
  - *etkili* (influential) *gözlem* ve
  - *yüksek etki* (high-leverage) *noktaları*

  birbiriyle iç içe girmiş ve çok yakın ilişkisi olan üç kavramdır.

62

- SAPAN GÖZLEM

  - *Sapan* terimi, bir veri setinde; diğer gözlemlerin yanında bazı özellikleri açısından tutarsız olan gözlem anlamına gelir.
  - Tahminlenen regresyon doğrusuna uyum göstermekte başarısız olan gözlemler sapan gözlem olarak değerlendirilebilirler.
  - Bir gözlem, beklenen sınırların dışındaki değerlere sahip bağımlı değişkenden ya da bağımsız değişken(ler)den dolayı sapan gözlem olabilir.

63

- SAPAN GÖZLEM

  - Sapan terimi, geri kalan örnek verilerinin yanında tutarsız olan bağımlı değişken değeri için kullanılmaktadır.
  - *Artıklardaki sapan* ifadesi ise gözlenen artıklar içinde beklenen kabul edilebilir değişkenlikten daha büyük değere sahip veri noktaları için kullanılır.
  - Genelde "sapan" *bağımlı değişkenin* ya da *artığın değeri* olarak kullanılmaktadır.

64

- SAPAN GÖZLEM İÇİN TEŞHİS YÖNTEMLERİ
  - Sapanlar, gözlenen artıkların analizi ile keşfedilebilir.
  - *Tanı İstatistikleri*; sapan gözlem, veri setindeki diğer gözlemlerle karşılaştırıldığında büyük değerde studendize artığa sahip gözlem olarak tanımlanabilir.
  - *Grafiksel Analiz*; Normal dağılım ve varyans heterojenliği durumlarını keşfetmek için plot edilen artıklar da sapanların tanımlanmasında etkili olurlar.

65

- YÜKSEK ETKİ NOKTALARI
  - Diğer veri noktalarından uzak olan bir nokta, regresyon sonuçları üzerinde hemen hemen diğer tüm noktalar kadar etkiye sahip olabilir. Bu tür gözlemlere *yüksek etki (levarage) noktaları* denir.
  - EKK sonuçları üzerinde bir veri noktasının olası etkisi, diğer noktalara göre **X**-uzayındaki, pozisyonu ile hesaplanır.
  - Genellikle bir veri noktası, **X**-uzayındaki veri noktalarının merkezinden uzaklaştıkça, regresyon sonuçlarını etkileme olasılığı da aynı oranda artar.

66

- YÜKSEK ETKİ NOKTALARI VE TEŞHİSİ
  - Bir yüksek etki noktası, veri setindeki diğer gözlemlerle karşılaştırıldığında büyük bir $h_{ii}$ değerine sahip gözlem olarak tanımlanabilir.
  - Etki kavramı tamamen bağımsız değişkenler ile ilgili olup, bağımlı değişkenlerle bir ilgisi yoktur.
  - **H** matrisinin köşegen elemanları, örnek noktaları ile örnek **X**-uzayının merkezi arasındaki öklid uzaklığının ölçüsüdür.

67

- EKK nın GİZLİ VARSAYIMI
  - Sıradan EKK yöntemi hesaplamalarda her gözleme eşit ağırlık vermektedir.
  - Bununla beraber, her gözlem çeşitli EKK sonuçlarında eşit etkiye sahip değildir.
  - Örneğin, basit doğrusal regresyon modelindeki eğim, bağımsız değişkenin ortalamadan en uzaktaki değerlerinden etkilenir.

NEDEN?

68

- ETKİLİ GÖZLEMLER
  - *Etkili gözlemler*, veri setindeki diğer gözlemlerle karşılaştırıldığında, bireysel olarak ya da birlikte uyumu yapılan regresyon denklemini aşırı ölçüde etkileyen gözlemlerdir.
  - Bu tanım görecelidir, araştırmacı *tanı istatistiklerini* kullanarak gözlemleri etki derecelerine göre sıralayabilir.

69

- ETKİ TİPLERİ
  - Bir gözlem tüm regresyon sonuçları üzerinde aynı etkiye sahip olmayabilir.
  - Bu gözlem regresyon sonuçlarından hangisini (veya hangilerini) etkilemektedir?
  - Bir gözlem,
    - tahminlenmiş regresyon katsayıları **b**,
    - tahminlenmiş **b** varyansları,
    - $\hat{\mathbf{y}}$ kestirilmiş değerleri ve/veya
    - uyum iyiliği istatistikleri üzerinde

    etkili olabilir.
  - Analizin ilk amacı hangi tip etkinin dikkate alınacağı sorusuna cevap aramaktır.

70

- ÜÇ KAVRAM ARASINDAKİ İLİŞKİ

  - Sapan bir gözlem, etkili gözlem olmayabilir,
  - Etkili bir gözlem, sapan gözlem olmayabilir,
  - Büyük artık değerine sahip gözlemlerin arzu edilmeyen bir durum olması küçük bir artığa karşılık gelen gözlemlerin ideal gözlem olduğu anlamına gelmemelidir.

  Neden1?

  - Sapan gözlemlerde olduğu gibi yüksek etki noktaları da etkili gözlem olmayabilir.
  - Buna karşın etkili gözlemlerin de yüksek etki noktası olması gerekli değildir.
  - Bununla birlikte yüksek etki noktaları potansiyel etkili gözlemler olarak değerlendirilebilirler.

71

- SAPANLARIN VE ETKİLİ NOKTALARIN KAYNAĞI

  - Bir çalışmada,
    - veriler toplanırken yapılan hatalardan veya
    - veri noktalarının değişik anakütlelerden gelmesinden

    dolayı bir veri noktası sapan ya da potansiyel etkili nokta olabilir.
  - Modelin süreci yeterince açıklayamadığı durumlarda, doğru bir nokta sapan gibi görünebilir.
  - Bu nokta için sapan artığına (outlier residual) sahiptir denir.

72

- NE YAPILMALI ?
  - Açık bir şekilde tanımlanabilen hatalar mümkünse düzeltilmelidir.
  - Mümkün değilse veri setinden çıkarılmalıdır.
  - Hatalı oldukları açık olarak tanımlanamayan ya da doğru oldukları belirlenen veri noktalarının, dikkatli bir şekilde incelenmeleri gereklidir.

73

- NE YAPILMAMALI ?
  - Modelin ya da tasarımın yetersizliklerini ortaya koyabilirler.
  - Sapan ve etkili gözlemlerin fark gözetmeksizin analizden atılması büyük bir hata olacaktır.
  - Bu özelliğe sahip veriler belki de araştırmada en fazla bilgi taşıyan gözlemlerdir.

74

- NEDEN ARTIKLAR ANALİZ EDİLİR?
  - Regresyon analizinde problem, $\boldsymbol{\varepsilon}$ nun gözlemlenemediği gibi aynı zamanda tahminlenememesinden kaynaklanmaktadır.
  - Hatalar modeldeki $\boldsymbol{\beta}$ vektörü bilinmediği için gözlemlenemezler.
  - Bununla birlikte **e** artığı, bir anlamda şans hatası $\boldsymbol{\varepsilon}$ u ölçmektedir.

75

# REGRESYON ANALİZİNDE HATALAR VE ARTIKLAR

- Hatalar;
  - o Sıfır ortalamalı,
  - o Sabit varyanslı,
  - o Birbirinden bağımsız
  - o Normal dağılış

  $\boldsymbol{\varepsilon} \sim N(\mathbf{0};\mathbf{I}\sigma^2)$

  gösterir.

- Artıklar;
  - o Sıfır ortalamalı,
  - o *Farklı varyanslı*,
  - o *Birbiri ile ilişkili*
  - o Normal dağılış

  $\mathbf{e} \sim N[\mathbf{0};(\mathbf{I}-\mathbf{H})\sigma^2]$

  gösterir.

76

- ARTIK VE HATALAR ARASINDAKİ FARKLAR
  - $\boldsymbol{\varepsilon}$ vektörünün elemanları birbirinden bağımsız ve aynı eşit varyansa sahip olsa bile,
    - **H** matrisi köşegen matris olmadıkça artıklar birbirinden bağımsız değildir,
    - **H** matrisinin köşegen elemanları birbirine eşit olmadıkça artıklar eşit varyanslı değildir.

  Not: **H** matrisi ve elemanlarının özellikleri için

  ☞ Ref. Tanım 40 ve 41

  Not: Artıklardaki farklı varyanslılık, her bir artığın standardize edilmesi ile düzeltilebilir.

77

- ARTIKLAR ARASINDAKİ KOVARYANSLAR
  - Eşitlikler, (I21.3) ve (I21.4)'den,
    - $V(\hat{y}_i)=\sigma^2 h_{ii}$ ve
    - $V(e_i)=\sigma^2(1- h_{ii})$

    olarak elde edilebilir.

    Not: Yorumlar için

    ☞ Ref. Tanım 42

  - $\sigma^2$ sabiti bir tarafa bırakılırsa, $h_{ii}$ değeri, $V(\hat{y}_i)$ ve $V(e_i)$'yi hesaplamaktadır.

78

- ARTIKLAR ARASINDAKİ KOVARYANSLAR
  - Eşitlik (I21.4)'den $e_i$ ve $e_j$ arasındaki kovaryans ,

    $$Cov\,(e_i, e_j) = -\sigma^2 h_{ij}$$

    olup, $e_i$ ve $e_j$ arasındaki korelasyon katsayısı

    $$Cor(e_i,e_j)=\rho_{ij}=\frac{-h_{ij}}{\sqrt{1-h_{ii}}\sqrt{1-h_{jj}}}$$

    olarak bulunabilir.
  - Eşitlikten görüldüğü gibi $e_i$ ve $e_j$ arasındaki korelasyon tamamen **H**'nin elemanları tarafından hesaplanmaktadır.

79

- HATA YERİNE ARTIĞIN KULLANILABİLMESİ İÇİN GEREKLİ KOŞULLAR
  - Eşitlik (3.13b) **ε** yerine **e**'nin kullanılabileceği göstermektedir.
  - Fakat bunun uygun bir yaklaşım olabilmesi için;
    - **H** matrisinin elemanları bazı özelliklere sahip olmalıdır,
    - Uyumu yapılan model doğru olmalıdır.

80

- NOTLAR
  - Bu iki durum gerçekte birbirinden bağımsız değildir.
  - Modelin yanlış belirlenmesinin **H** matrisi üzerinde etki oluşturacağı unutulmamalıdır.

81

- HATA YERİNE ARTIĞIN KULLANILABİLMESİ İÇİN GEREKLİ KOŞULLAR
  - **ε** yerine ***e*** nin uygun bir şekilde kullanılabilmesi için;
    - **X** matrisinin sıralarının homojen olması, bu durumda **H** matrisinin köşegen elemanları yaklaşık olarak eşittir, ve
    - **H** matrisinin köşegen dışı elemanlarının yeterince küçük olması gereklidir.

82

- AMAÇ NEDİR?
  - Artıkların analizinde amaç **e** değerlerinin incelenerek **ε** ile ilgili varsayımlarda ortaya çıkan eksikliklerin yorumlanmasıdır.
  - Bazı problemlerde, modelin yanlış kurulması, **ε** değerleri ile **e** değerleri arasında uygun bir ilişki kurulabilmesini engeller.
  - Bu gibi durumlarda artıklarda gözlemlenen bazı belirtiler, modelin uygun olmadığının ve/veya varsayımların gerçekleşmediğinin belirtisi olabilir.

83

- ARTIK ANALİZ TİPLERİ
  - Artıklar üzerine yapılan inceleme iki genel kısımda toplanabilir;
    - Tek bir sapan gözlemin testi için oluşturulan artık testleri (istatistikleri),
    - Artıklar üzerine oluşturulan grafiksel tanı yöntemleri.
  - Özellikle **H** matrisine olan bağımlılık nedeni ile, tanı amacıyla kullanılacak artıkların dönüştürülmüş tiplerinin kullanılması tercih edilmektedir.
  - Artıkların dönüşümünde amaç, ölçek parametresinden bağımsız bir dağılıma sahip artıkların elde edilmesidir.

84

- ALTERNATİF YÖNTEM

- Alternatif olarak, artıklar seçilmiş bir kovaryans yapısına sahip olacak şekilde dönüştürülebilirler.
- Bu durumda *n-p* elemanlı ve elemanları ilişkisiz olan bir artık vektörünün elde edilmesi ile ilgilenilir.



- DÖNÜŞTÜRÜLMÜŞ ARTIK İSTATİSTİKLERİ

  - Sıradan artıklar, her bir artığın varyansı hem $\sigma$ hem de $h_{ii}$ değerlerinin bir fonksiyonu olduğu için ölçeği bağımlı bir dağılıma sahiptirler.
  - Bu değerlere bağımlı olmayan bir artık tipinin regresyon tanısı olarak kullanılması daha faydalıdır.
  - Bu amaçla $e_i$ yerine:

$$f(e_i, \sigma_i) = \frac{e_i}{\sigma_i} \qquad \textbf{(6.1)}$$

  tanı istatistiği kullanılabilir.

  ☞ Tanım 37

- DÖNÜŞTÜRÜLMÜŞ ARTIK İSTATİSTİKLERİ
  - Burada $\sigma_i$, $i$-inci artığın standart sapmasıdır. Diğer bir deyişle $\sigma^2(\mathbf{e})$ matrisinin $i$-inci köşegen elemanının kare köküdür.
  - Eşitlik (6.1) in dört özel durumu mevcuttur:
    - o normalize artık,
    - o standardize artık,
    - o içsel studendize artık,
    - ☞ Tanım 38
    - o dışsal studendize artık.
    - ☞ Tanım 39

87

- NORMALİZE ARTIK
  - Eşitlik (6.1) de $\sigma_i$ yerine $(\mathbf{e}^T\mathbf{e})^{1/2}$ konularak,

$$a_i = \frac{e_i}{\sqrt{\mathbf{e}^T\mathbf{e}}} \qquad i=1,2,...,n \qquad \textbf{(6.2}a\textbf{)}$$

  - Elde edilecek olan **a** artık vektörü birim vektöre dönüşmektedir.
  - $max\ |a_i|$ istatistiği maksimum normlanmış artık olarak adlandırılır.

88

- STANDARDİZE ARTIK

  - Eşitlik (6.1) de $\sigma_i$ yerine,

$$s = \sqrt{\frac{\mathbf{e}^T\mathbf{e}}{n-p}}$$

konularak,

$$b_i = \frac{e_i}{s} \qquad \textbf{(6.2}b\textbf{)}$$

elde edilir.

89

- İÇSEL STUDENDİZE ARTIK

  - Eşitlik (6.1) de $\sigma_i$ yerine,

$$\sigma_i = s\sqrt{1-h_{ii}} \qquad \textbf{(6.2}c\textbf{)}$$

olarak alınırsa, *i*-inci içsel studendize artık,

$$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}} \qquad \textbf{(6.2}d\textbf{)}$$

olarak tanımlanmıştır.

  - Artık varyanslarının önemli ölçüde değişim gösterdiği durumlar için, $e_i/s$ yerine $r_i$ istatistiğinin kullanılması önerilmiştir.

90

- DIŞSAL STUDENDİZE ARTIK
  - Eşitlik (6.1) de $\sigma_i$ yerine,

$$\sigma_i = s_{(i)}\sqrt{1-h_{ii}} \qquad \textbf{(6.2}e\textbf{)}$$

alınarak,

$$r_i^* = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}} = \frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}_{(i)}}{s_{(i)}\sqrt{1-h_{ii}}} \qquad \textbf{(6.2}f\textbf{)}$$

özdeşliği ile verilir.
  - Burada $y_i - \mathbf{x}_i^T\boldsymbol{\beta}_{(i)}$ uyumunda $y_i$ değeri kullanılmamıştır.
  - $r_i^*$ istatistiği jackknife artık olarak adlandırılmıştır.

91

- DIŞSAL STUDENDİZE ARTIK
  - $r_i^*$ değerleri birbirinden bağımsız değildir.
  - Bu ifadedeki,

$$s_{(i)}^2 = \frac{\mathbf{y}_{(i)}^T\left(\mathbf{I} - \mathbf{H}_{(i)}\right)\mathbf{y}_{(i)}}{n-p-1} \qquad \textbf{(6.3)}$$

$i$-inci gözlem ihmal edildiğinde elde edilen artığın ortalama karesel hatasının tahminidir ve
  - $\mathbf{H}_{(i)} = \mathbf{X}_{(i)}\left(\mathbf{X}_{(i)}^T\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}_{(i)}^T \qquad \textbf{(6.4)}$

matrisi $i$-inci gözlemin silindiği $\mathbf{X}_{(i)}$ matrisi için izdüşüm matrisidir.

92

- DÖRT İSTATİSTİK ARASINDAKİ İLİŞKİ

- Bu amaçla $s_{(i)}$ ifadesini $s$ ye göre yazmak gereklidir:

$$s_{(i)}^2 = s^2\left(\frac{n-p-r_i^2}{n-p-1}\right) \qquad \textbf{(6.5)}$$

- Bu eşitlik kullanılarak,

$$b_i = a_i\sqrt{n-p} \qquad \textbf{(6.6}a\textbf{)}$$

$$r_i = \frac{b_i}{\sqrt{1-h_{ii}}} = a_i\sqrt{\frac{n-p}{1-h_{ii}}} \qquad \textbf{(6.6}b\textbf{)}$$

$$r_i^* = \frac{a_i\sqrt{n-p-1}}{\sqrt{(1-h_{ii})-a_i^2}} \qquad \textbf{(6.6}c\textbf{)}$$

$$r_i^* = r_i\sqrt{\frac{n-p-1}{n-p-r_i^2}} = \frac{r_i}{\left(\frac{n-p-r_i^2}{n-p-1}\right)^{1/2}} \qquad \textbf{(6.6}d\textbf{)}$$

eşitlikleri elde edilir,

93

- DÖRT İSTATİSTİKTEN HANGİSİ KULLANILIR?

  - Dört istatistiğin hepsinde de artıkların birbirinden bağımsız dağılış göstermediği görülmektedir.
  - Büyük örnek hacimleri söz konusu olduğunda, pek çok problem için bu olumsuz özellik ihmal edilebilir.
  - Tanı amacı ile ele alındıklarında normalize artıkların ve standardize artıklar $e_i$'nin varyansını ortaya çıkarmazlar.
  - Bu nedenle model yetersizliğinin ve sapan gözlemlerin araştırılmasında tercih edilen istatistikler $r_i$ ya da $r_i^*$'dır.

94

- NİÇİN $r_i$ YA DA $r_i^*$ ?

  - Veri setinde sapan gözlemin bulunmadığı durumlarda, $s^2$ ve $s^2_{(i)}$ nin her ikisi de $\sigma^2$'nin sapmasız tahminini sağlar.
  - Buna karşın veri setinde bir sapan gözlem bulunması durumunda, örneğin *j*-inci gözlem, $s^2_{(j)}$ hala sapmasızdır ve $r_j^*$ merkezi olmayan bir *t*-dağılımına sahiptir.
  - Kalan $r_i^*$ ve tüm $r_i$ istatistikleri $\sigma^2$'nin sapmalı tahminleyicisi olacaklardır.

95

- NİÇİN $r_i$ YA DA $r_i^*$ ?

  - Bu nedenle dışsal studendize artığın kullanılması veri setinde tek bir sapan gözlemin bulunduğu durumlarda daha uygun sonuçlar verir.
  - **H** matrisinin köşegen elemanlarının (dolayısı ile $e_i$'nin varyansının) değişken bir yapıya sahip oldukları durumlarda $r_i$'nin kullanılması tavsiye edilir.

96

- DIŞSAL STUDENDİZE ARTIĞIN AVANTAJLARI
  - $r_i^*$, ($n$-$p$-1) serbestlik dereceli $t$-dağılımı gösterdiği için artıkların aşırı değere sahip olup olmadıklarının değerlendirilmesinde uygun bir kriterdir, çünkü tabloları mevcuttur.
  - Eşitlik (6.6$d$) den görüldüğü gibi $r_i^*$ istatistiği $r_i$'nin monotonik fakat doğrusal olmayan bir transformasyonudur ve $r_i^2 \rightarrow (n\text{-}p)$ iken $r_i^{*2} \rightarrow \infty$ olduğundan, $r_i^*$ büyük sapmaları $r_i$ istatistiğine göre daha iyi yansıtmaktadır.
  - $s_{(i)}$ tahmini, $i$-inci gözlemdeki toplam hata problemine karşı duyarsızdır.
  - $r_i^*$ istatistiği eşit varyansa sahiptir.

97

- TEK BİR SAPAN İÇİN TEST YÖNTEMLERİ
  - Gözlenmiş artıkların davranışlarının tam testleri mevcut değildir.
  - Yaklaşık olarak ya da göreceli değerlendirmeler kullanılmak zorundadır.
  - Standardize ya da studendize artıkların bir sapan gözlemin kontrolü amacı ile kullanılması bir çoklu test prosedürüdür.
  - Çünkü bu test prosedürü anlamlılık seviyelerini bulmak için birinci *Bonferroni eşitsizliği* üzerine oluşturulmaktadır.

98

- TEK BİR SAPAN İÇİN TEST YÖNTEMLERİ
  - Test edilecek artık, $n$ örnek hacmi içindeki en büyük artık olabilecektir ve $\alpha$ olasılığı için uygun değerler belirlenmelidir.
  - $\alpha$ olasılığına göre birinci dereceden Bonferroni sınırı için $\alpha=\alpha^*/n$ olmak üzere $t$-kritik değerinin kullanılması önerilmektedir.
  - Burada $\alpha^*$ istenilen kapsamlı anlamlılık seviyesidir.
  - Sapan gözlemlerin belirlenmesi için Bonferroni anlamlılık seviyeleri artıklar arasındaki korelasyonun aşırı derecede büyük olmadığı durumlar için uygun bir güvenceye sahiptir.

99

- TEST YÖNTEMLERİNİN GENEL YAPISI
  - Konu ile ilgili birkaç prosedür mevcuttur.
  - Bu prosedürler genellikle, veri setinde tek bir sapan gözlem bulunduğu varsayımına dayanmaktadırlar.
  - Bu prosedürlerin genel yapısı,
    - $T(x_i, y_i)$ istatistikleri için,

    $Pr\{T(x_i, y_i)>C_\alpha$ / en fazla bir sapan mevcut$\}\leq\alpha$ şeklindeki $C_\alpha$ değerinin bulunması,
    - Eğer $T(x_i y_i)> C_\alpha$ ise

    $i$-inci gözlemin sapan olarak belirlenmesi, adımlarından oluşur.
  - Bir sapan gözlem için aday durum bilinmediğinde test genellikle maksimum $r_i$ ya da $r_i^*$ değeri üzerine oluşturulur.

100

- TIETJEN, MOORE VE BECKMAN PROSEDÜRÜ
  - Basit regresyon modelini incelemişler, bu model için,

    $$T(x_i y_i)=R_n= max\,|\,r_i\,| \equiv r_{max} \quad \textbf{(6.7a)}$$

    eşitliğini önermişlerdir.
  - Basit regresyon için kritik değerler simulasyon ile elde edilmiştir.
  - Büyük bir $R_n$ değeri veri setinde sapan bir gözlemin varlığını belirtir.

101

- TIETJEN, MOORE VE BECKMAN PROSEDÜRÜ
  - $R_n$ istatistiğinin bileşenleri, $e_i/s(1\text{-}h_{ii})^{1/2}$, incelendiğinde paydanın da sıfır olabileceği durumlar olduğu görülmektedir.

    Nasıl 1?
  - Diğer bir deyişle paydanın sıfır olduğu durumlarda $e_i$ daima sıfır değerini almaktadır.
  - Bu nedenle $e_i$=0 olduğunda $R_n$=0 olarak tanımlanmıştır.

102

- PRESCOTT (1975) PROSEDÜRÜ
  - Genel doğrusal model için;

$$R_n^* = \max\left|\frac{e_i}{\bar{s}}\right|$$

istatistiğini önerilmiştir.
  - Burada $\bar{s}^2$, artıkların tahminlenmiş ortalama varyansıdır.
  - Artıkların ortalama varyansı $(n\text{-}p)\sigma^2/n$ şeklindedir.
  - Bu durumda, $\bar{s}^2=(n\text{-}p)s^2/n$ olduğu görülebilir.
  - Sonuç olarak,

$$R_n^* = n^{1/2}\ max|a_i| \qquad \textbf{(6.7}b\textbf{)}$$

istatistiği kullanılabilir.

☞ İspat 36

103

- PRESCOTT (1975) PROSEDÜRÜ
  - $|a| = max\ |a_i|$ için 100(1-α)'lık yüzde noktasının üst sınırı;

$U=[n(n\text{-}p)aF/(n(n\text{-}p\text{-}1+F))]^{1/2}$

şeklinde tanımlamıştır.

Not: Burada $F$, serbestlik derecesi 1 ve $(n\text{-}p\text{-}1)$ olan $F$ dağılımı için %(1-α) noktasındaki değeri tanımlamaktadır.
  - Eşitlik (6.7*a*) ve (6.7*b*) için yaklaşık kritik değerleri,

$$\sqrt{\frac{(n-p)F}{n-p-1+F}} \qquad \textbf{(6.7}c\textbf{)}$$

şeklindedir.

Not: Kritik değerler, her bir $p$ değeri, $n$ örnek hacmi ve α olasılık değerleri için Tablo 6.1 de verilmiştir.

104

- BONFERRONİ TESTİ
  - Her bir dışsal studentized artık, ($n$-$p$-1) serbestlik dereceli $t$-dağılışı gösterir.
  - Bonferroni kritik değeri;

    $$t(1-\alpha/2n; n-p-1)$$

    olarak tanımlanmıştır.
  - ☞ Ref. Örnek 6.2

105

- GRAFİKSEL TANI YÖNTEMLERİ
  - Artıkların oluşturdukları şekillerin düzeni, büyüklüklerinden çok daha fazla bilgi taşır.
  - Araştırma için gerekli bilginin elde edilebilmesi için genellikle birden fazla farklı plotun analizinin gerekli olabileceği unutulmamalıdır.

106

- ARTIK PLOTLARI TİPLERİ
    - Artıkların frekans dağılımı; histogramlar, nokta grafikler vb.
    - Artıkların normal ya da yarı normal plotları.
    - Uyumu yapılmış değerlere karşı artık plotları.
    - Bağımsız değişkene karşı artık plotları.
    - Eklenmiş değişken plotları.
    - Kısmi regresyon etki plotları.
    - Kısmi artık plotları.
    - Zamana karşı artık plotları.
    - Artıkların bir gecikmeli plotu

107

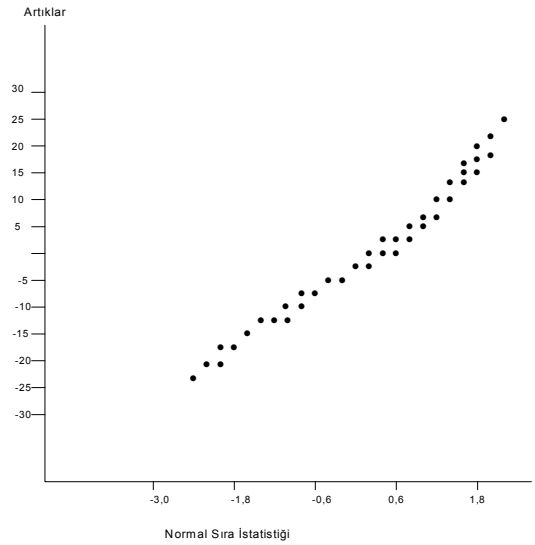
- ARTIKLARIN FREKANS DAĞILIMI
    - Artıkların basit bir frekans dağılımı, çarpıklık, birden fazla mod, basıklık vb. gibi normallik varsayımını geçersiz kılabilecek durumların ortaya çıkarılmasında oldukça faydalıdır.
    - Bununla birlikte normallik varsayımının kontrolü için bu tip plotların kullanımında sonuçların güvenilir olabilmesi için örnek hacminin büyük olması gereklidir.
    - Gözlem sayısının az olması durumunda nokta grafikler kullanılabilir.
        - ☞ Ref. Örnek 6.3

108

- NORMAL OLASILIK PLOTLARI
  - Uygun örnek hacmi için normal sıra (order) istatistiklerine karşı sıralanmış artıkların plotudur.
  - Normal sıra istatistikleri, ortalaması sıfır varyansı bir olan normal dağılımdan gelen dizine sokulmuş gözlemlerin beklenen değeridir.
  - Gözlenmiş ve küçükten büyüğe düzenlenmiş artıkların normal sıra istatistiklerine karşı plot edilmesi normal plotları verir.
  - Artıklar normal dağılımdan alınan bir örneği temsil ediyorsa normal plotun beklenen sonucu, orijinden geçen ve eğimi artıkların standart sapmasına bağlı olan düz bir doğru şeklinde olmasıdır, (bkz. Şekil 6.1).

109

# Şekil 6.1

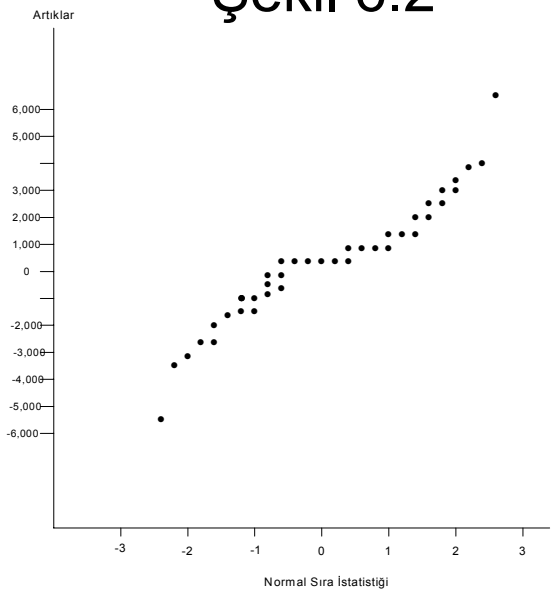



110

- NORMAL OLASILIK PLOTLARI
  - Örnek sıra istatistiklerinin örnekleme değişkenliğine bağlı olarak düz doğrudan rassal bazı sapmalar olabilecektir.
  - Bir normal plotta beklenen düz doğrudan ayrılışlar normal dağılımdan faklılığı belirtir.
  - Bir çarpık dağılım, normal plotun bir eğri şeklinde oluşmasına neden olur.
  - Bu eğrinin yönü çarpıklığın yönüne bağlıdır.
  - S-şeklindeki bir eğri, eğrinin yönüne bağlı kalın veya ince kuyruklu bir dağılımı belirtir, (bkz. Şekil 6.2).

111

# Şekil 6.2




112

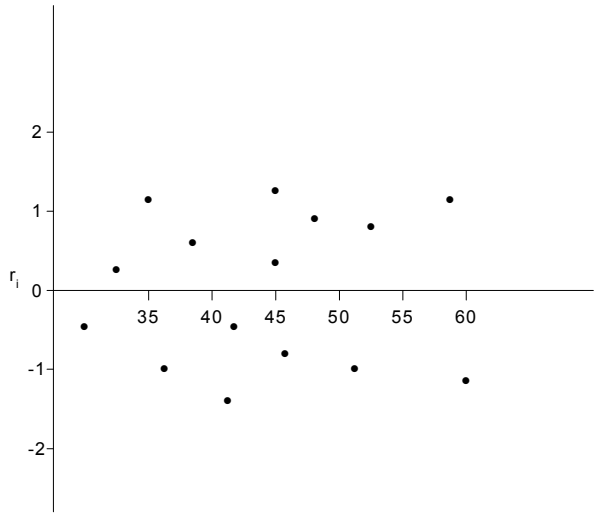
- NORMAL OLASILIK PLOTLARI
  - Kalın kuyruklu dağılımlar, normal dağılıma göre daha fazla ekstrem gözlemlere sahiptir.
  - İnce kuyruklu dağılımlar ise daha az ekstrem gözlem içerir.
  - Bu arada dikkat edilmesi gereken bir konu da, modelde mevcut bazı kusurlarında normal dağılıştan farklılık etkisine benzer durumlar gösterebileceğidir.
  - Örneğin, farklı varyanslılık veya sapan artıklar ince kuyruklu bir dağılım görüntüsü verebilir.
  - Örneklemedeki değişkenlik nedeniyle, normal dağılımdan farklılık büyük olmadıkça, küçük örnekler için normal olasılık plotları fazla bilgilendirici olmayabilir.

113

- UYUMU YAPILMIŞ DEĞERLERE KARŞI ARTIK PLOTLARI
  - Artıkların uyumu yapılmış değerlere ya da **X** matrisinin elemanlarına karşı oluşturulan plotu,
    - $r_i$ için yaklaşık ortogonal,
    - $e_i$ için ise tam ortogonaldir.
  - varsayımlar doğru ise $r=0$ doğrusunun çevresinde dağılan artıkların yaklaşık olarak $r=\pm 2$ sınırlarının içinde olması beklenir, (Şekil 6.3).

114

# Şekil 6.3

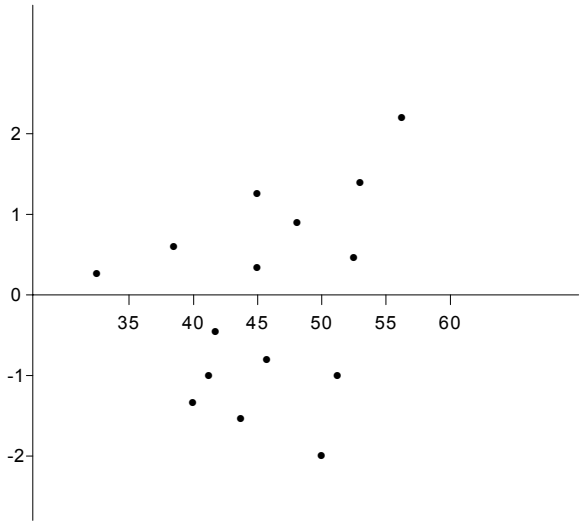



115

- FARKLI VARYANS DURUMU
  - $\hat{y}_i$ değerlerine göre giderek genişleyen bir yapı göstermesi farklı varyanslılığın belirtisidir, (Şekil 6.4).
  - Bağımlı değişkenin varyansı ortalamasının bir fonksiyonu $\sigma^2=f(\mu)$ ise bu tip durumlarla karşılaşılması beklenir.
  - Ayrıca hatalar eklenebilir değilse, örneğin çarpımsal ise, Şekil 6.4 de gösterilen durumla karşılaşılabilir.

116

# Şekil 6.4




117

- ASİMETRİNİN BELİRTİLERİ
  - Artıkların dağılımındaki herhangi bir asimetri modelde ya da temel varsayımlarda bir problem olduğunu belirtir.
  - Örneğin, negatif artıkların değer olarak küçük fakat sayıca fazla, pozitif artıkların değer olarak büyük fakat sayıca daha az olması, simetrik bir dağılış göstermesi gereken artıkların pozitif yönde çarpık dağılım gösterdiklerinin göstergesidir.
  - Çarpık dağılımın bu tip bir plot ile teşhis edilmesi, artıkların normal plotu ya da histogramına göre daha basittir.

118

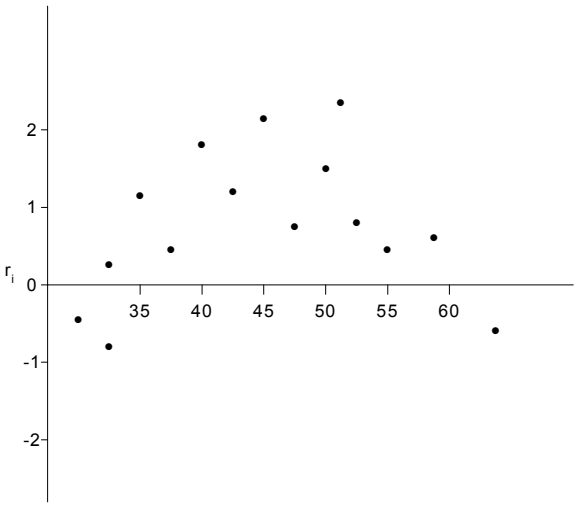
- MODEL YETERSİZLİĞİ-I

  - $\hat{y}_i$'nin bazı bölgelerinde negatif diğer bölgelerinde ise pozitif artıkların fazla olması
    - verilerdeki sistematik bir hatayı ya da
    - modelde önemli bir değişkenin ihmal edildiğini ya da
    - mevcut değişkenlerden birisinin karesel formunun modele ilave edilmesi

    gerektiğini belirtir, (Şekil 6.5).
  - Artıkların büyük bir çoğunluğunun içinde bulunduğu bandın dışında bulunan artık (veya artıklar) sapan gözlem olarak nitelendirilir.

119

# Şekil 6.5

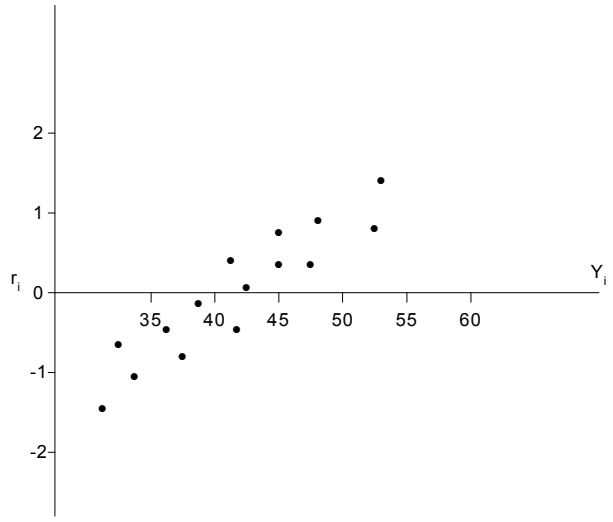



120

- MODEL YETERSİZLİĞİ-II
  - Artıklar yükselen veya aşağıya inen bir düz bant şeklinde görüntü çizebilirler, (Şekil 6.6).
  - Bu tip durumlarla genellikle sabit terimin modelde bulunmaması (unutulması) halinde karşılaşılır.

121

# Şekil 6.6




122

- BAĞIMSIZ DEĞİŞKENE KARŞI ARTIK PLOTU
  - Çok değişkenli regresyonda, $p$-boyutlu bir uzayı (bağımsız değişken uzayı) iki boyutlu bir plot ile temsil etmek, yatay eksenin seçiminde problem oluşturmaktadır.
  - İlk aşamada $p$-boyutlu uzaydaki bir vektör seçilir ve veri noktaları bu tek vektör üzerine izdüşümlenir.
  - $x_1$ e karşı $r_i$'nin plotunda, $x_2$ ihmal edilerek $x_1$ in aynı değerleri aynı absis değerine sahiptir.
  - Buna karşın $r_i$'nin $\hat{y}_i$'ye karşı oluşturulan plotunda aynı absis değeri sadece aynı kestirim değerli durumlar için geçerlidir.

123

- BAĞIMSIZ DEĞİŞKENE KARŞI ARTIK PLOTU
  - Bağımsız değişkene karşı plotlar, sadece ilgili bağımsız değişken açısından modelde bir yetersizlik bulunup bulunmadığını ortaya koyabilirler.
  - $x_1$ için oluşturulan bir plot $x_1^2$ teriminin ilave edilmesi gerektiğini ya da $V(\varepsilon_i)= x_{1i}\sigma^2$ şeklinde bir farklı varyanslılığı tanımlayabilir.
  - Bu plotlar $x_1$ ve $x_2$ arasındaki etkileşim etkilerini belirleme yeteneğine sahip değildir.

124

- BAĞIMSIZ DEĞİŞKENE KARŞI ARTIK PLOTU
  - Artıkların bir bağımsız değişkene karşı oluşturulan plotları, $\hat{y}_i$ değerlerine karşı oluşturulan plotlara benzer şekilde yorumlanır.
  - Sıfır çevresinde giderek büyüyen bir dağılım farklı varyanslılığı belirtir.
  - Bu tip plotların kullanılmasıyla bağımsız değişkenin yüksek dereceden polinomlarının modele dahil edilip edilmemesi gerekliliği ortaya çıkacaktır.

125

- BAĞIMSIZ DEĞİŞKENE KARŞI ARTIK PLOTU
  - Bağımsız değişkenin yüksek dereceli teriminin modelde oluşturduğu olumsuzluğun, modeldeki diğer bağımsız değişkenlerin dağılımı ve etkileri nedeniyle anlaşılması güç olabilecektir.
  - Modelde birden fazla bağımsız değişenin bulunması durumunda kısmi regresyon etki plotlarının kullanılması daha uygundur.
  - Yüksek etki noktaları ve özellikle iki veya daha fazla bağımsız değişkenin kombinasyonlarının oluşturduğu etkili gözlemlerin tek değişkenli plotlarla belirlenmesi oldukça güçtür.

126

- ARTIKLARIN ZAMANA KARŞI PLOTLARI
  - Zaman sırası dikkatle alınarak belirli periyotlarda alınan gözlemler genelde otokorelasyonlu artıklara sahip olabilirler.
  - Bu tür gözlemlerin artıkları önceki artıkların polinomları olarak ifade edilebilir.
  - Otokorelasyon, artıkların birbirini takip edecek şekilde sıfır etrafında sistematik yapı oluşturmasıdır.
  - Otokorelasyonun belirlenmesi için en çok kullanılan yöntemlerden biri işaret (run) testidir.

127

- ARTIKLARIN ZAMANA KARŞI PLOTLARI İÇİN TESTLER
  - Test birbirini takip eden pozitif ve negatif artıklardaki dizin sayısını dikkate almaktadır.
  - Daha sonra bu dizin sayısı, artıkların birbirinden bağımsız olduğu sıfır hipotezi altında ortaya çıkması beklenen dizin sayısı ile karşılaştırılır.
    - ☞ Ref. Örnek 6.4
  - Daha az dizin sayısı pozitif otokorelasyondan ileri gelebilirken artıklardaki çok sayıdaki dizin sayısı negatif otokorelasyonun bir göstergesi olabilir.

128

- ARTIKLARIN ZAMANA KARŞI PLOTLARI İÇİN TESTLER

- Eğer $n_1$ ve $n_2$ gözlem sayıları 10'dan fazla ise dizin dağılımı için normal dağılışa yaklaşım kullanılabilir.
  - Dağılım ortalaması,

$$\mu = \frac{2n_1n_2}{n_1+n_2} \qquad \textbf{(6.8}a\textbf{)}$$

  - varyansı,

$$\sigma^2 = \frac{2n_1n_2\left(2n_1n_2-n_1-n_2\right)}{\left(n_1+n_2\right)^2\left(n_1+n_2-1\right)} \qquad \textbf{(6.8}b\textbf{)}$$

  - standart normal sapması,

$$z = \frac{\left(u-\mu+1/2\right)}{\sigma} \qquad \textbf{(6.8}c\textbf{)}$$

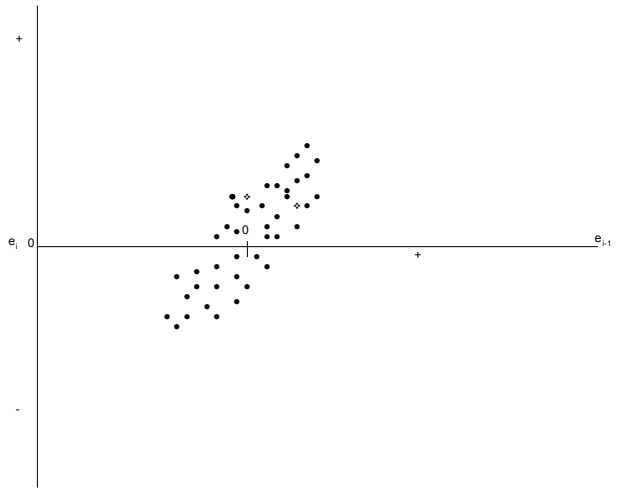
Not: Eşitlik (6.8*c*)'deki (1/2) değeri süreklilik için düzeltme faktörü olarak kullanılmaktadır.

129

- ARTIKLARIN BİR GECİKMELİ PLOTU
  - Zaman serisi verilerindeki otokorelasyon, her bir artığın kendinden bir önceki artığa karşı plotu oluşturarak daha açık bir şekilde ortaya çıkarılabilir.
  - Verilerdeki bir pozitif otokorelasyon, Şekil 6.7 de olduğu gibi pozitif eğimli bir noktalar seti oluşturur.

130

# Şekil 6.7




131

- ARTIKLARIN BİR GECİKMELİ PLOTU İÇİN TEST
  - Artıklarda mevcut olan otokrelasyon Durbin-Watson testi ile de belirlenebilir.
  - Durbin-Watson test istatistiği,

$$d = \frac{\sum_{i=2}^{n}\left(e_i - e_{i-1}\right)^2}{\sum_{i=1}^{n} e_i^2} \quad \textbf{(6.9)}$$

  - Durbin-Watson testi kullanılırken anlamlı bir sonuç elde edilmesi için verilerin zamana göre düzenlenmesi (eğer zaman serisi verisi değilse) gereklidir.

132

- YÜKSEK ETKİ NOKTALARI
  - Sadece artıkların incelenmesi yüksek etki noktalarının belirlenmesi için yeterli olmayabilir.

    Neden 2?
  - Bu kısımda bir noktanın etkisinin ölçümü ile ilgili dört kriter incelenecektir.
    - İz düşüm matrisinin *i*-inci köşegen elemanı.
    - Mahalanobis uzaklığı
    - Ağırlıklı karesel standardize uzaklık (WSSD).
    - Genişletilmiş $\mathbf{H}_{xy}$ matrisinin *i*-inci elemanı.

133

- YÜKSEK ETKİ NOKTALARI
  - Ayrıca yüksek etki değerleri ile artıklar L-R plotu adı verilen bir tek grafiksel gösterimde birlikte ele alınacaklardır.
  - Bu plot yüksek etki noktaları ile sapan gözlemlerin ayrıştırılmasına olanak verir.

134

- İZ DÜŞÜM MATRİSİNİN KÖŞEGEN ELEMANLARI
  - İzdüşüm matrisinin elemanları eşitlik (6.10*a*) köşegen elemanları eşitlik (6.10*b*) ile tanımlanmıştır.

$$h_{ii} = \mathbf{x}_i^T \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{x}_i \qquad \textbf{(6.10}\boldsymbol{a}\textbf{)}$$

$$h_{ij} = \mathbf{x}_i^T \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{x}_j \qquad \textbf{(6.10}\boldsymbol{b}\textbf{)}$$

  - Bu değerler, artıkların büyüklüğü, aralarındaki varyans, kovaryans yapıları ve uyumu yapılmış değerlerin belirlenmesinde önemli rol oynarlar.
  - Bu nedenle hem $r_i^*$ hem de $h_{ii}$ değerlerinin birlikte incelenmesi önerilmektedir,
    - sapanların belirlenmesi için $r_i^*$,
    - yüksek etki noktalarının belirlenmesi için $h_{ii}$.

135

- $h_{ii}$ İÇİN KRİTİK DEĞERLER-I
  - Ortaya çıkan soru hangi $h_{ii}$ değerinin büyük olarak kabul edileceğidir.
  - $h_{ii}$ için genel kullanıma sahip üç kritik değer tanımlanmıştır.
    - Huber (1981),

$$h_{ii} > 0.2 \qquad \textbf{(6.11}\boldsymbol{a}\textbf{)}$$

      noktalarının yüksek etki noktası olarak sınıflanmasını önermiştir.

      Not:Kestirilmiş değerlerin beş ya da daha az eşdeğer gözlem tarafından belirlendiği gözlemler için özel bir dikkatin gösterilmesi zorunludur.

136

- $h_{ii}$ İÇİN KRİTİK DEĞERLER-II
  - Hoaglin ve Welsch (1978),

$$h_{ii} > \frac{2p}{n} \quad \textbf{(6.11}b\textbf{)}$$

noktalarını yüksek etki noktaları olarak tanımlamıştır.

137

- $h_{ii}$ İÇİN KRİTİK DEĞERLER-III
  - Model sabit bir terim içeriyorsa ve **X** matrisinin sıraları birbirinden bağımsız olarak $N_{p\text{-}1}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ şeklinde dağılıyorsa, Chatterjee ve Hadi, (1988).

$$\frac{n-p}{p-1}\frac{h_{ii}-(1/n)}{1-h_{ii}} \sim F(p\text{-}1,\ n\text{-}p)$$

olup,

$$h_{ii} \geq \frac{nF(p-1)+(n-p)}{nF(p-1)+(n(n-p))} \quad \textbf{(6.11}c\textbf{)}$$

noktaları yüksek etki noktasıdır. Not:Burada $F$ serbestlik dereceleri ($p$-1, $n$-$p$) olan %100(1-$\alpha$)'lık $F$ dağılımının tablo değeridir.

☞ Ref. Tanım 43

138

- KRİTİK DEĞERLER İLE İLGİLİ YORUMLAR.
  - **X** matrisinin sıraları normal dağılım gösterseler dahi eşitlik (6.11*c*) ile tanımlanan kritik değer pek fazla kullanıma sahip değildir.
  - Yukarıda $h_{ii}$ için tanımlanan üç kriter mekanik olarak kullanılmamalıdır.
  - **H** matrisinin tüm köşegen elemanlarının birbiri ile karşılaştırılması tavsiye edilir.
  - Böyle bir karşılaştırmanın en iyi yolu $h_{ii}$ değerlerinin indeks plotu, ağaç yaprak ve/veya kutu plotları şeklindeki grafiksel gösterimidir.
    - ☞ Ref. Örnek 6.4 ve 6.5

139

- MAHALANOBİS UZAKLIĞI
  - Aşağıdaki istatistik
    $M_i^2=(\mathbf{o}_i-\bar{\mathbf{o}})\mathbf{C}^{-1}(\mathbf{o}_i-\bar{\mathbf{o}})^T$ **(6.12*a*)**
    *i*-inci durumun karesel Mahalanobis uzaklığı olarak adlandırılır.
    - ☞ Ref. Tanım 44
  - SAS ve SPSS gibi paket programlar bu istatistiği hesaplamaktadırlar.
  - $M_i^2$ değerleri belirli bir anlamlılık seviyesinde $p$-2 serbestlik dereceli $\chi^2$ dağılımı tablo değerleri ile karşılaştırılabilir.
  - Karesel Mahalanobis uzaklığının, iz düşüm matrisinin köşegen elemanları ile ilişkisi,
    $M_i^2=(n-1)[h_{ii}-(1/n)]$ **(6.12*b*)**
    olarak tanımlanmıştır.

140

- AĞIRLIKLI KARESEL STANDARDİZE UZAKLIK
  - Daniel ve Wood (1980) ağırlıklı karesel standardize uzaklığın,

$$WSSD_i = \frac{\sum_{j=1}^{p} c_{ij}^2}{s_y^2}, \; i = 1,\ldots,n \qquad \textbf{(6.13}a\textbf{)}$$

    kullanılmasını önermişlerdir.
    ☞ Ref. Tanım 45
  - Burada,

$$s_y^2 = \frac{\sum_{j=1}^{n} \left( y_j - \overline{y} \right)^2}{n-1} \qquad \textbf{(6.13}b\textbf{)}$$

141

- AĞIRLIKLI KARESEL STANDARDİZE UZAKLIK
  - $WSSD_i$, $j$-inci değişken için tahminlenmiş regresyon katsayısının büyüklüğünü, bu değişkenin göreceli önemini belirten bir ağırlık olarak kullanarak, $j$-inci değişkenin ortalamasından $\overline{x}_j$, $x_{ij}$ değerlerinin karesel uzaklığının toplamını veren bir ölçümdür.
  - Eğer $i$-inci gözlem en az bir değişkene göre extrem durumu tanımlıyorsa (ki onun tahminlenmiş katsayısı büyüktür) $WSSD_i$ değeri büyük olacaktır.
    ☞ Ref. Örnek 6.5 ve 6.6

142

- GENİŞLETİLMİŞ İZ DÜŞÜM MATRİSİNİN KÖŞEGEN ELEMANLARI
  - Bir tanı ölçümü olarak tek başına $h_{ii}$ değerinin kullanılmasının dezavantajı **y** vektöründe içerilen bilgiyi ihmal etmesidir.
  - Bu eksikliği gidermek amacıyla **X** matrisi **y** vektörüyle genişletilsin **Z**=(**X**:**y**).
  - **Z** matrisi hem **X** hem de **y**'deki bilgiyi içerdiği için ona ait olan izdüşüm matrisinin $\mathbf{H}_z$ köşegen elemanlarının $h_{zii}$ kullanımı daha uygun olabilir.

143

- GENİŞLETİLMİŞ İZ DÜŞÜM MATRİSİNİN KÖŞEGEN ELEMANLARI
  - Eşitlik (I37.1) den,
    $$h_{zii}=\mathbf{z}_i^T(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{z}_i=h_{ii}+(e_i^2/\mathbf{e}^T\mathbf{e}) \quad \textbf{(6.14)}$$
    yazılabilir.
    - ☞ Ref. İspat 37
  - $h_{ii}$ ya da $e_i^2$ (ya da her ikisinin) büyük olması durumunda $h_{zii}$ büyük olacaktır.
  - Fakat $h_{zii}$, **X**-uzayındaki yüksek etki noktaları ile **Z**-uzayındaki sapan gözlemleri birbirinden ayıramayacaktır.
    - ☞ Ref. Örnek 6.5 ve 6.6

144

- L-R PLOTU
  - Yüksek etki noktaları ile sapan gözlemleri birbirinden ayıran etkin bir plot, L-R plot olarak adlandırılır.
  - Bu plot etkili değerleri ve artıkları tek bir plot olarak birleştirir.
  - Bu yöntemde iz düşüm matrisinin köşegen elemanları karesel normalize artıklara $a_i^2$ karşı plot edilir.
  - L-R plot aşağıdaki şartları sağlamak zorundadır:
    - $0 \le h_{ii} \le 1$
    - $0 \le a_i^2 \le 1$
    - $h_{ii} + a_i^2 \le 1$
      - ☞ Ref. Örnek 6.5 ve 6.6

145

- YORUMLAR
  - Buraya kadar yapılan açıklamalardan ve uygulamalardan görüldüğü gibi ne sapan gözlemin ne de yüksek etki noktalarının etkili gözlem olmaları gerekli değildir.
  - Etkili gözlemlerin belirlenmesi için ilave prosedürlere ihtiyaç vardır.

146

- ETKİLİ GÖZLEMLER
  - Farklı amaçlar için kullanılabilecek tanı istatistiklerinden dört tanesi aşağıda kullanım amaçları ile birlikte özetlenmiştir:

    1. **b** üzerindeki etkiyi ölçen Cook'un $D_i$'si ,

    2. $\hat{y}_i$ üzerindeki etkiyi ölçen $DFFITS_i$ ,

    3. $\mathbf{b}_j$ üzerindeki etkiyi ölçen $DFBETAS_{j\,(i)}$ ,

    4. Parametre tahminlerinin varyans–kovaryans matrisi üzerindeki etkiyi ölçen $COVRATIO_i$ .

147

- COOK UZAKLIĞI
  - Cook'un tanımlamış olduğu ölçüm $D_i$ ile belirtilmiştir.
  - Bir gözlem ihmal edildiğinde **b** vektöründeki değişmeyi ölçmektedir.
  - Bu istatistik, ihmal edilen gözlemin tüm regresyon katsayıları üzerindeki etkisinin ölçümünü sağlar.
  - Normallik varsayımı altında, **b** vektörü için oluşturulan $\boldsymbol{\beta}$'nın %100(1-$\alpha$)'lık ortak güven bölgesi,

    $$\frac{(\boldsymbol{\beta}-\mathbf{b})^T(\mathbf{X}^T\mathbf{X})(\boldsymbol{\beta}-\mathbf{b})}{ps^2} \le F_{(1-\alpha,p,n-p)} \quad \textbf{(6.15}\boldsymbol{a}\textbf{)}$$

    şeklinde merkezi **b** vektörü olan bir elipsoit tanımlar.

148

- TANI İSTATİSTİĞİ
  - Bu elipsoidin eşyükselti eğrileri ($\mathbf{X}^T\mathbf{X}$) matrisinin özdeğer ve özvektörleri ile tanımlanır.
  - Cook (1977) tarafından önerilen;

$$D_i = \frac{\left(\mathbf{b}_{(i)} - \mathbf{b}\right)^T \left(\mathbf{X}^T\mathbf{X}\right)\left(\mathbf{b}_{(i)} - \mathbf{b}\right)}{ps^2} \qquad \textbf{(6.15}b\textbf{)}$$

  etki ölçüm istatistiği elde edilir.
  - Bu ölçüm, sabitlenen $\mathbf{X}^T\mathbf{X}$'e göre $\mathbf{b}$'den $\mathbf{b}_{(i)}$'ye olan uzaklığın karesini verir.
  - Diğer bir ifade ile *i*-inci gözlemin etkisi, bu gözlem silindiğinde (6.15*b*) ile tanımlanan güven elipsoidinin merkezinde ya da eşdeğer olarak tahminlenmiş katsayılar üzerinde oluşan değişim ile ölçümlenebilir.

149

- TANI İSTATİSTİĞİ
  - Cook un hesaplamalarda kolaylık sağlamak amacıyla önerdiği formül:

$$D_i = \left[\frac{y_i - \mathbf{x}_i^T\mathbf{b}}{s\sqrt{1-h_{ii}}}\right]^2 \frac{h_{ii}}{p(1-h_{ii})} = \frac{r_i^2}{p}\frac{V(\hat{y}_i)}{V(e_i)}$$

$$D_i = \frac{1}{p} r_i^2 \frac{h_{ii}}{1-h_{ii}} \qquad \textbf{(6.15}c\textbf{)}$$

  - $D_i$, iki ölçümü bir araya getirmektedir:
    - İlk bileşen $r_i$, istatistiğin şansa bağlı bileşeni olup sapan gözlemler ile ilgili bilgiyi taşır,
    - ikinci bileşen $h_{ii}$ (ya da $h_{ii}/(1- h_{ii})$), potansiyel yüksek-etki noktaları ile ilgili bilgi vermektedir

150

- TANI İSTATİSTİĞİ ÜZERİNE YORUMLAR

- $D_i$, iki ölçümü bir araya getirmektedir:
  - İlk bileşen $r_i$, istatistiğin şansa bağlı bileşeni olup sapan gözlemler ile ilgili bilgiyi taşır,
  - ikinci bileşen $h_{ii}$ (ya da $h_{ii}/(1- h_{ii})$), potansiyel yüksek-etki noktaları ile ilgili bilgi vermektedir.
- Eğer $h_{ii}$ ve $r_i^2$ büyük ise $D_i$ değeri de büyük olacaktır.
- Bu durumda $i$-inci gözlem etkili gözlem olarak değerlendirilebilir.
- Hangi büyüklükteki $D_i$ değerinin bir gözlemin etkili olduğunu belirttiği sorusunun cevabı gerçekte, karşılaşılan probleme bağlıdır.
- Cook, $r_i$ ve $V(\hat{y}_i)/V(e_i)$ değerlerinin ayrı ayrı incelenmesi ile ek bilginin elde edilebileceğini belirtmiştir.

151

- TANI İSTATİSTİĞİ İÇİN KRİTİK DEĞERLER

- $D_i$ değerleri, $F(p, n-p)$ dağılımı için hesaplanmış değerler yardımı ile bir olasılık ölçeğine dönüştürülebilirler.
- Bununla birlikte $D_i$ nin dağılımı bir $F$-dağılımı değildir.
- Bu karşılaştırma sadece $D_i$'yi bilinen bir ölçeğe dönüştürebilmek için kullanılır.
- Bu olay bir anlamlılık testi olarak algılanmamalıdır.
  - ☞ Ref. Tanım 47

152

- TANI İSTATİSTİĞİ İÇİN KRİTİK DEĞERLER
  - 1 den büyük $D_i$ değerleri **b** ve $\mathbf{b}_{(i)}$ arasındaki uzaklık için %50'lik bir güven elipsoidi tanımlar.
  - Kritik değerler:
    - $D_i$ <0,10, $D_i$ <0,20 ise *i*-inci gözlemin tahmin değerleri üzerinde biraz etkisi vardır.
    - $D_i$ >0,50 ise *i*-inci gözlem fonksiyonun tahmini üzerinde büyük bir etkiye sahiptir.
      - ☞ Ref. Örnek 6.7

153

- **WELSCH-KUH UZAKLIĞI ($DFFITS_i$)**
  - Bir gözlemin silinmesinin öngörümleme üzerinde oluşturduğu etkinin ölçümlenmesi için tanı istatistiği; Belshley, Kuh, ve Welsch 1980;
    $DFFIT_i=\hat{\mathbf{y}}_i-\hat{\mathbf{y}}_{i(i)}=\mathbf{x}_i[\mathbf{b}-\mathbf{b}_{(i)}]=h_{ii}e_i/(1-h_{ii})$ **(6.16*a*)**
  - Kestirilmiş değer $\hat{y}_i$ üzerinde *i*-inci gözlemin etkisi, uyumun $\hat{y}_i=\mathbf{x}_i\mathbf{b}$ standart hatasına göre göreceli olarak ölçümlenebilir.

$$DFFITS_i=WK_i=\left|r_i^*\right|\sqrt{\frac{h_{ii}}{1-h_{ii}}} \quad \textbf{(6.16}b\textbf{)}$$

154

- DFFITS$_i$ ve D$_i$ ARASINDAKİ İLİŞKİ
  - *DFFTIS* $\hat{y}_i$ ve $\hat{y}_{i(i)}$ arasındaki farkı ölçeklendirir.
  - $\hat{y}_{i(i)}$, *i*-inci gözlem için tahminlenen ortalamadır; ancak *i*-inci gözlem, $\beta$'nın tahminlenmesinde kullanılmamıştır.
  - *DFFITS*$_i$ (*WK*$_i$) ile Cook'un $D_i$ istatistiği arasındaki ilişki,

$$D_i = (WK_i)^2 \left( \frac{s^2_{(i)}}{ps^2} \right) \quad \textbf{(6.16}c\textbf{)}$$

155

- TANI İSTATİSTİĞİ İÇİN KRİTİK DEĞERLER
  - *WK*$_i$ nin büyük değerleri *i*-inci gözlemin uyum üzerinde etkili olduğunu belirtir.
  - *DFFITS*, mutlak değer olarak $2(p/n)^{1/2}$'den büyük olduğunda etkili gözlem durumu söz konusudur.
  - *DFFITS*$_i$ ile Cook'un $D_i$'si arasındaki ilişkiden $D_i$ için kritik değer $4/n$ olarak alınabilir.
    Not: $s^2$ ve $s^2_{(i)}$ arasındaki fark ihmal edilmiştir.
  - ☞ Ref. Örnek 6.7

156

- TANI İSTATİSTİĞİ İÇİN KRİTİK DEĞERLER
  - $WK_i$ bir $t$-dağılımı göstermez. Bununla birlikte $t$-istatistiğine benzer bir davranış yapısı mevcuttur.
  - $r_i^* \sim t_{(n-p-1)}$ olduğu için kritik değer olarak Chatterjee ve Hadi (1988),

    $$t\sqrt{p/(n-p)}$$

    değerini önermiştir .

    Not:Burada $t$ değeri tablodan $t_{(n-p-1)}$ olarak bulunabilir.

157

- BİR REGRESYON KATSAYISI ÜZERİNDEKİ ETKİ (DFBETAS)
  - Regresyon parametre tahminleri üzerinde oluşan etki için tanımlanan en basit istatistik Belsley, Kuh ve Welsch (1980) tarafından *DFBETA* olarak adlandırılan istatistiktir:

    $$DFBETA_i = \mathbf{b} - \mathbf{b}_{(i)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i \frac{e_i}{1-h_{ii}} \qquad \textbf{(6.17}a\textbf{)}$$

  - Bu istatistik parametre tahminlerinin varyansına göre ölçeklenmemiştir.
  - Ölçekleme işlemi uygulanan istatistik *DFBETAS* olarak adlandırılmıştır.

    $$DFBETAS_{j(i)} = r_i^* \frac{w_{ij}}{\sqrt{\mathbf{w}_j^T\mathbf{w}_j}} \frac{1}{\sqrt{1-h_{ii}}} \textbf{(6.17}b\textbf{)}$$

  - Tek tek, regresyon katsayıları için etkili gözlemler $DFBETAS_{j(i)}$ ile tanımlanabilir.

158

- DFBETAS

  - Her bir $DFBETAS_{j(i)}$, $i$-inci gözlem silindiğinde $b_j$'deki standardize değişmeyi verir. Alternatif olarak,

$$DFBETAS_{j(i)} = \frac{b_j - b_{j(i)}}{s_{(i)}\sqrt{c_{jj}}} \qquad \textbf{(6.17}c\textbf{)}$$

    Not:Burada $c_{jj}$, $(\mathbf{X}^T\mathbf{X})^{-1}$ matrisinin $(j+1)$'inci köşegen elemanıdır.
  - $DFBETAS_{j(i)}$, $b_j$'deki değişiklikleri, standart hatasının katları cinsinden ölçer.

159

- DFBETAS İÇİN KRİTİK DEĞERLER

  - Bu istatistik bir $t$-istatistiği olarak görünmesine rağmen, $t$-istatistiğine dayalı bir anlamlılık testi olarak yorumlanmamalıdır.
  - 2 den büyük olan $DFBETAS_{j(i)}$ değerleri büyük bir değişmeye işaret eder.
  - Ancak bu durum genelde olası değildir. Belsley ,Kuh ve Welsch (1980) tarafından tavsiye edilen kritik nokta $2/\sqrt{n}$'dir.

160

- VARYANS ORANI (COVRATİO)

- **b** ve $\mathbf{b}_{(i)}$ için tahminlenmiş varyanslar karşılaştırılarak *i*-inci gözlemin etkisi değerlendirilebilir.
- *i*-inci gözlemin, tahminlenen regresyon katsayılarının varyans-kovaryans matrisi üzerindeki etkisi, iki varyans-kovaryans matrisinin determinantlarının oranı ile ölçülebilir.
- Belsley, Kuh ve Welsch (1980), determinantların oranının,

$$VR_i = \frac{\left|s^2_{(i)}\left(\mathbf{X}^T_{(i)}\mathbf{X}_{(i)}\right)^{-1}\right|}{\left|s^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right|}$$

$$= \left(\frac{s^2_{(i)}}{s^2}\right)^p \frac{\left|\mathbf{X}^T\mathbf{X}\right|}{\left|\mathbf{X}^T_{(i)}\mathbf{X}_{(i)}\right|} \qquad \textbf{(6.18}a\textbf{)}$$

$$COVRATIO_i = VR_i = \left(\frac{n-p-r_i^2}{n-p-1}\right)^p \left(\frac{1}{1-h_{ii}}\right) \qquad \textbf{(6.18}b\textbf{)}$$

161

- COVRATIO
  - Varyans-kovaryans matrisinin determinantı varyansın genelleştirilmiş ölçümüdür.
  - Bu nedenle *COVRATIO*, *i*-inci gözlemin, regresyon katsayıları tahmininin doğruluğu üzerindeki etkisini yansıtır.

162

- COVRATIO İÇİN KRİTİK DEĞERLER

  - Değerlerin 1 civarında olması, *i*-inci gözlemin, tahminlerin doğruluğu üzerindeki etkisinin az olduğunu gösterir.
  - *COVRATIO* değerinin 1'den büyük olması, *i*-inci gözlemin varlığının tahminlerin doğruluğunu arttırdığı anlamına gelmektedir;
  - Oran 1 den küçük olduğunda ise gözlemin varlığı, tahminlerin doğruluğunu azaltıyor demektir.

  ☞ Ref. Tanım 49

163

- COVRATIO İÇİN KRİTİK DEĞERLER

  - Tüm gözlemlerin kovaryans matrisi üzerinde eşit etkiye sahip olduğu ideal durum için $VR_i$ değeri yaklaşık olarak bir değerine sahiptir.
  - Bir değerinden oluşan sapmalar *i*-inci gözlemin potansiyel etkili olduğunu belirtir.
  - Belsley, Kuh ve Welsch (1980), $VR_i$ için iki ekstrem durum;
    - $|r_i| \geq 2$, $h_{ii}=1/n$ durumu
    - $h_{ii} \geq 2p/n$, $r_i=0$ durumu

    için yaklaşık olarak kritik değerleri belirlemişlerdir.

164

- COVRATIO İÇİN KRİTİK DEĞERLER
  - İlk durum için kritik değer, yaklaşık olarak; $VR_i \leq 1+3p/(n-p)$
  - ikinci durum için kritik değer yaklaşık olarak, $VR_i \geq 1+3p/(n-p)$

165

- FVARATIO
  - Bir gözlemin silinmesi durumunda $\hat{y}_i$ değerinin varyansında oluşan değişim ile ilgilenildiğinde;

$$FVARATIO_i = \frac{V\left[\hat{y}_{i(i)}\right]}{V\left(\hat{y}_i\right)} = \frac{s^2_{(i)}}{s^2\left(1-h_{ii}\right)} \quad \textbf{(6.19)}$$

    elde edilir.
    - ☞ Ref. Tanım 50
  - Bu eşitlik $s^2_{(i)}/s^2$ oranının $p$-inci kuvveti hariç *COVRATIO*'ya benzerlik gösterir.

166

- DFTSTAT
  - Normal dağılım varsayımı geçerli olduğunda, *i*-inci gözlemin silinmesinin regresyon katsayılarının *t*-istatistiği üzerindeki etkisi;

$$DFTSTAT_{ij} = \frac{b_j}{s\sqrt{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{jj}}} - \frac{b_{j(i)}}{s_{(i)}\sqrt{\left(\mathbf{X}^T_{(i)}\mathbf{X}_{(i)}\right)^{-1}_{jj}}} \quad \textbf{(6.20)}$$

    tanı istatistiği ile incelenir.
  - Tanı istatistiğinin incelenmesi, ilgili gözlemin hipotez testi sonuçlarını etkileyip etkilemediğini gösterir.

167

## KRİTİK DEĞERLER İÇİN YORUMLAR

- Kritik noktaların belirlenmesi için literatürde tanımlanan üç bilgi kaynağı mevcuttur:
  - Dışsal ölçekleme,
  - içsel ölçekleme ve
  - farklılıklar (gaps) yaklaşımı.
- Bu yaklaşımlardan ilk ikisi açıklanmıştır.

168

- DIŞSAL ÖLÇEKLEME
  - Dışsal ölçeklemede kritik değerler istatistik teorisi kullanılarak belirlenir.
  - Normal dağılış varsayımı altında, ilgili tanı istatistiğinden bağımsız olarak tahminlenmiş uygun bir standart hata ile ölçeklenmiş olan, $r_i^*$, *DFBETAS* ve *DFFITS* gibi *t*-değişkeni benzeri dağılışı gösteren tanılar için mutlak değerce iki değerinden büyük olan değerler önemli olarak kabul edilirler.
  - $r_i^*$ istatistiğinin örnek hacmine doğrudan bağımlılığı diğerlerine göre daha az olduğundan bu prosedürün kritik noktaların belirlenmesinde en faydalı olduğu istatistik $r_i^*$'dır.

169

- DIŞSAL ÖLÇEKLEME
  - Mutlak kritik değerler örnek hacmine doğrudan bağımlı olan *DFBETAS* ve *DFFITS* tanıları için de ekstrem değerlerin belirlenmesinde kullanılabilirler.
  - Buna karşın *COVRATIO* ya da $h_{ii}$ için bu tanılara ait bir standart hata bulunmadığından, mutlak kritik değerler kullanılamaz.
  - Örnek hacmini de dikkate alan kritik değerler, örnek hacmi düzeltmeli kritik değerler olarak adlandırılırlar
  - Bu değerler *DFBETAS*, *DFFITS*, $h_{ii}$ ve *COVRATIO* için ilgili kısımlarda tanımlanmışlardır.

170

- DIŞSAL ÖLÇEKLEME
  - Hem mutlak hem de örnek hacmi düzeltmeli kritik değerler arasındaki ilişki özellikle büyük veri setleri için önemlidir.
  - Büyük veri setlerinde her hangi bir gözlemin silinmesi ile $|DFBETAS|$ ya da $|DFFITS|$ değerlerinde büyük değişiklikler oluşmaz.
  - Diğer bir deyişle $n$ değerinin büyük olduğu durumlarda mutlak anlamda etkili olan her hangi bir gözlemin bulunması olası değildir.
  - Örnek hacmi düzeltmeli kritik değerler bu amaçla kullanılabilir.

171

- İÇSEL ÖLÇEKLEME
  - Bu yaklaşım araştırmadan elde edilen tanı sonuçları serisindeki göreceli olarak büyük olan değerleri tanımlar.
  - Gözlemlerin sırayla çıkarılması sonucunda $n$ adet tanı değeri elde edilir.
  - İzdüşüm matrisi köşegen elemanları ve *DFFIT* gibi ya da $p$ adet *DFBETA* tanıları bir veri seti için tanı serisi oluşturmaktadır.

172

- İÇSEL ÖLÇEKLEME
  - Tukey ile tanımlanan kartiller arası değişim aralığı (interquartile range) *š* her bir veri seti için hesaplanır
  - (7/2)š değerini aşanlar etkili gözlem olarak değerlendirilir.
  - Eğer tanılar normal dağılış gösteriyorsa etkili olmayan bir gözlemin bu değeri aşma olasılığı %0.1'den daha azdır.
  - Kartiller arası değişim aralığı istatistiği, tanı serisi değerlerinin normalden farklı bir dağılıma sahip olduğu durumlar için standart sapmadan daha güçlü (robust) bir yayılım ölçüsü tahmini sağlar.

173

- ÇOKLU DOĞRUSAL BAĞLANTI TEŞHİS YÖNTEMLERİ
  - Çoklu doğrusal bağlantı teşhis yöntemleri iki kısımda incelenebilir:
    - İnformal Teşhis Yöntemleri
    - Formal Teşhis Yöntemleri

174

- İNFORMAL TEŞHİS YÖNTEMLERİ
  - Açıklayıcı bir değişken modele eklendiğinde ya da modelden çıkarıldığında; bir gözlem değiştirildiğinde ya da silindiğinde regresyon katsayılarında büyük değişikliklerin meydana gelmesi
  - Modeldeki önemli açıklayıcı değişkenlerin regresyon katsayılarının anlamsız bulunması
  - Modeldeki regresyon katsayılarının işaretlerinin beklentilerimizle uyum göstermemesi
  - Basit korelasyon katsayılarının büyük çıkması
  - Önemli bir açıklayıcı değişkeni temsil eden regresyon katsayısının güven aralığının çok geniş bulunması

175

- İNFORMAL TEŞHİS YÖNTEMLERİNİN DEZAVANTAJLARI
  - Bu yöntemler bize;
    - kantitatif bilgiler sağlamaz ve
    - çoklu doğrusal bağlantının yapısı hakkında bilgi vermez.
  - Bu nedenle formal teşhis yöntemlerine başvurulur.

176

- FORMAL TEŞHİS YÖNTEMLERİ
  - Regresyon değişkenleri arasındaki basit korelasyonlardır. Bu yöntem çoklu doğrusal bağlantının derecesi hakkında bize bilgi vermez.
  - Şartlanma Sayısı, $\mathbf{X}^T\mathbf{X}$ korelasyon matrisi özdeğerler sistemi
  - Varyans Artış Faktörü (Variance Inflation Factor-VIF)

177

- ŞARTLANMA SAYISI
  - Bir kare matris küçük determinant değerine sahipse o matrisin kötü şartlandığı ifade edilir.
  - Matris kare değilse $\mathbf{X}^T\mathbf{X}$ matrisinin determinant değeri incelenir ve küçük ise kötü şartlanmıştır.
  - Bununla birlikte küçük bir determinant bir matrisin tersinin alınabilirliği ile ilgili her şeyi ortaya koymaz.
  - Amaç $\mathbf{A}^{-1}$ matrisinin elde edilebilirliğinin araştırılması olduğunda, kötü şartlanma teriminin $\mathbf{A}^{-1}$ ile ilişkisinin ortaya konması gerekli olacaktır.

178

- ŞARTLANMA SAYISI
  - **X** matrisinin şartlanma sayısı bu matrisin tekil değerleri ile ve dolayısıyla $\mathbf{X}^T\mathbf{X}$ matrisinin özdeğerleri ile ilişkilidir.
  - Rankı $r$ olan herhangi $n \times p$ boyutlu **X** matrisi verilmiş olsun.
  - **X** matrisinin tekil değerleri ve $\mathbf{X}^T\mathbf{X}$ matrisinin özdeğerleri $\mu_1 \geq \mu_2 \geq ... \geq \mu_p$ ve $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ şeklinde düzenlenebilir.
  - Şartlanma sayısı,

$$\kappa(\mathbf{X}) = \frac{\mu_1}{\mu_p} = \sqrt{\frac{\lambda_1}{\lambda_p}} \geq 1 \qquad \textbf{(6.21}a\textbf{)}$$

179

- ŞARTLANMA SAYISI İÇİN KRİTİK DEĞER
  - Şartlanma sayısı $\kappa(\mathbf{X})$ için minimum sınır bir değeridir.
  - Bu değeri **X** matrisinin sütunları ortogonal olduğunda alır.
  - $\kappa(\mathbf{X})$'nın büyük değerleri **X** matrisinin doğrusal bağlantıya sahip olduğunu belirtir.
  - Şartlanma sayısı büyük olan bir **X** matrisi için kötü şartlandığı söylenir.
  - Hangi büyüklükteki bir şartlanma indisinin verilerin doğrusal bağımlılığa sahip olduğunu ifade eden bir baz mevcut değildir.
  - Bu deneysel olarak belirlenmesi gereken bir sorundur.
  - Belsley, Kuh ve Welsch (1980), 5 ile 10 arasındaki şartlanma indisinin zayıf bağımlılığı, 30 ile 100 arasındaki değerlerin orta ve güçlü bağımlılığı tanımladığını belirtmişlerdir.

180

- ŞARTLANMA SAYISININ PROBLEMLERİ

- Eşitlik (6.21*a*) ile tanımlanmış olan $\kappa(\mathbf{X})$ ile ilgili iki problem aşağıda tanımlanmıştır:
  - $\kappa(\mathbf{X})$ sütun ölçeklemesine karşı duyarlıdır
  - $\mathbf{X}$ uzayındaki bir veya birkaç ekstrem noktadan aşırı derecede etkilenir.
- İlk problem iki farklı şekilde aşılmaya çalışılır.
  - Birincisi, $\mathbf{X}$ matrisinin sütunları eşit uzunlukta olacak şekilde normalize edilir. Bu uzunluk genellikle 1 olacak şekilde seçilir.
  - İkincisi, eğer model sabit bir terim içeriyorsa $\mathbf{X}$ matrisinin her bir sütunu ortalaması sıfır varyansı bir olacak şekilde standardize edilir.
- İkinci problem için ise tanı yöntemleri kullanılır?

181

- ŞARTLANMA İNDEKSİ

- $\mathbf{X}$ matrisinin sütunları arasındaki her bir tam doğrusal bağımlılığa karşılık tekil değerlerden biri sıfır değerini alır.
- Yaklaşık doğrusal bağımlılık için bu özellik genellendiğinde yaklaşık doğrusal bağımlılığın var olması durumunda küçük bir tekil değer (ya da özdeğer) ortaya çıkar.
- Kötü şartlanmanın derecesi en küçük tekil değerin en büyük tekil değerden ne kadar küçük olduğuna bağlıdır.
- Diğer bir deyişle $\mu_{max}$ ölçümlenebilecek küçüklüğe karşı bir baz oluşturmaktadır.

182

- ŞARTLANMA İNDEKSİ
  - Şartlanma indeksi

$$\eta_k = \frac{\mu_{\max}}{\mu_k} \quad k=1,2,\ldots\ldots p \quad \textbf{(6.21}b\textbf{)}$$

  - Eşitlik (6.21*b*) $n \times p$ boyutundaki **X** matrisi için $k$-ıncı şartlanma indeksini tanımlamaktadır.
  - Eşitlikten tüm $k$'lar için $\eta_k \geq 1$ olduğu görülmektedir.
  - $\eta_k$ için en büyük değer verilen matris için şartlanma sayısını tanımlayacaktır.
  - Büyük şartlanma indisi değerleri, matrisin sütunları arasındaki kuvvetli bağımlılıkları tanımlar.

183

- VARYANS ARTIŞ FAKTÖRÜ (VIF)
  - VIF, regresyon modelinde $X$' ler arasındaki ilişkinin etkisini ölçer.
  - Açıklayıcı değişkenler arasındaki hangi ilişkinin, tahminlerin kesinliğini hangi derecede azalttığını açıklar.
  - VIF değerleri, tahminlenmiş regresyon katsayılarının varyans-kovaryans matrisinden yola çıkarak hesaplanabilir.
  - Çoklu doğrusal bağlantının etkisini ölçmek için aşağıda belirtilen standartlaştırılmış regresyon modelinden yararlanılır.

$$y_i' = \beta_1' x_{i1}' + \ldots + \beta_{p-1}' x_{i\,p-1}' + \varepsilon_i \quad \textbf{(6.22)}$$

184

- VARYANS ARTIŞ FAKTÖRÜ (VIF) İÇİN FORMÜLLER

- Burada dönüştürülmüş değişkenler ile orijinal değişkenler arasındaki ilişki

$$y_i' = \frac{1}{\sqrt{n-1}}\left(\frac{y_i - \bar{y}}{s_y}\right) \quad \textbf{(6.23}a\textbf{)}$$

$$x_{ik}' = \frac{1}{\sqrt{n-1}}\left(\frac{x_{ik} - \bar{x}_k}{s_k}\right) \quad \textbf{(6.23}b\textbf{)}$$

- Burada kullanılan standart sapmalar,

$$s_Y = \sqrt{\frac{\sum\left(y_i - \bar{y}\right)^2}{n-1}} \quad \textbf{(6.23}c\textbf{)}$$

$$s_k = \sqrt{\frac{\sum\left(x_{ik} - \bar{x}_k\right)^2}{n-1}} \quad \left(k = 1, 2, \ldots, p-1\right)\textbf{(6.23}d\textbf{)}$$

- Burada dönüştürülmüş parametreler ile orijinal parametreler arasındaki ilişki

$$\beta_k' = \frac{\beta_k s_k}{s_Y} \quad \textbf{(6.23}e\textbf{)}$$

185

- VIF DEĞERİNİN ELDE EDİLİŞİ
  - Standartlaştırılmış regresyon modeli tahmin edildiğinde $b_k'$ standartlaştırılmış katsayılar elde edilir.
  - Açıklayıcı değişkenlerin varyans-kovaryans matrisini $\left[\mathbf{X}^T\mathbf{X}\right] = \mathbf{R}_{XX}$ kullanarak $\sigma^2\left\{\mathbf{b}'\right\} = \left(\sigma'\right)^2\mathbf{R}_{XX}^{-1}$ katsayıların varyansını hesaplanır.
  - $\left(\sigma'\right)^2$ dönüştürülmüş modelin varyansıdır.

186

- VIF FORMÜLLERİ
  - Varyans artış faktörü VIF değerleri, varyans-kovaryans ters matrisinin köşegeninde yer alan değerlerdir.

    $$VIF_i = r_{ii}^{-1} \qquad \textbf{(6.24}a\textbf{)}$$

  - Ayrıca VIF değerleri,

    $$VIF = \frac{1}{1 - R_k^{\,2}} \qquad k = 1, 2, ..., p-1 \quad \textbf{(6.24}b\textbf{)}$$

  - Böylece $b_k'$ ların varyansı

    $\sigma^2\{b_k\} = (\sigma')^2 (VIF)_k$ **(6.24*c*)**

    şeklinde bulunur.

187

- VARYANS ARTIŞ FAKTÖRÜ (VIF)
  - $R_k$, $x_k$ değişkenini diğer (*p-2*) adet *x* değişkeni üzerine regrese ettiğimizde elde edilen açıklanan varyanstır.
  - Eğer $X_k$ değişkeni diğer (*p-2*) *X* değişkeni ile yüksek derecede ilişkili ise $R_k^{\,2}$ ve dolayısıyla VIF da büyük olacaktır.
  - Bu durum $b_k'$' ların varyansını arttırır ve testlerin yapılmasını zorlaştırır.

188

- **VIF HANGİ DEĞERİ İÇİN ÇOKLU DOĞRUSAL BAĞLANTI PROBLEMİ OLUŞTURUR?**
  - $x$ değişkenleri arasındaki büyük VIF değerleri çoklu doğrusal bağlantının kuvvetini belirlemede kullanılır.
    - Değer olarak 10' u geçen açıklayıcı değişkenler arasındaki en büyük VIF değeri, çoklu doğrusal bağlantının en küçük kareler tahminlerini etkilediğini gösterir.
    - Ayrıca VIF değerlerinin ortalaması $\left(\overline{VIF}\right)$, standartlaştırılmış regresyon katsayıları $b_k^{'}$' ların, gerçek $\beta_k^{'}$' lardan ne kadar uzakta olduğunu belirlememizde bilgi sağlar.
      - ☞ Ref. Tanım 51

189

- SONUÇ
  - Çoklu Doğrusal Bağlantı modelde güçlü ise,
    - katsayılar doğru tahminlenemez,
    - katsayılara ait standart hatalar büyük çıkar ve buna bağlı olarak ta yanlış yorumlar ve kestirimler yapılır.
      - ☞ Ref. Örnek 6.8

190

- DOĞRUSAL BAĞLANTI NOKTALARININ TEŞHİSİ

- *Şartlanma sayısı üzerindeki etki*: **X** matrisinin şartlanma sayısı üzerinde, *i*-nci sıranın etkisini değerlendirmek amacıyla önerilen ölçüm;

$$K_i = \frac{\left|\tilde{\kappa}_{(i)} - \kappa\right|}{\kappa} \qquad \textbf{(6.25)}$$

- Burada κ, eşitlik (6.21*a*) ve $\tilde{\kappa}_{(i)}$ eşitlik (T52.3)'den elde edilmektedir.
- $K_i$ istatistiği, **X** matrisinin *i*-inci sırasının silinmesinin şartlanma sayısında oluşturduğu göreceli değişikliği tanımlamaktadır.
- Eğer $K_i$ büyük ve pozitif ise $\mathbf{x}_i$ sırasının silinmesi şartlanma sayısını büyütecektir.
- $K_i$ büyük ve negatif ise $\mathbf{x}_i$ sırasının silinmesi şartlanma sayısını azaltacaktır.

191