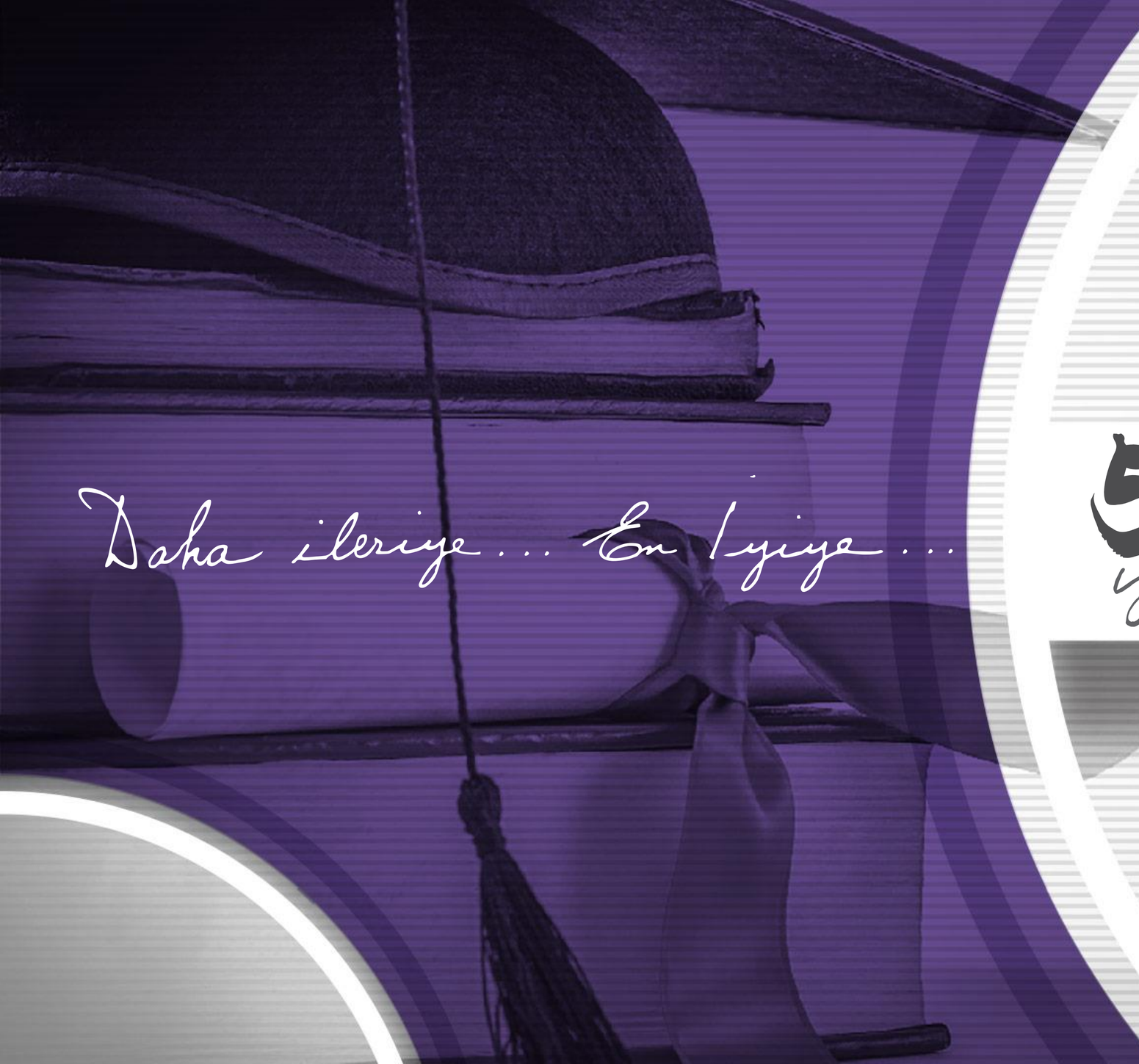


Daha ileriye... En İyiye...

50. yıl

HACETTEPE ÜNİVERSİTESİ

www.**hacettepe**.edu.tr

ÇOKLU REGRESYON

- Birden fazla bağımsız değişken/kestirici değişken kullanarak bir bağımlı değişkenin değerlerini kestirmeye çalıştığımız regresyon analizidir.
- Y: bağımlı değişken
- X1, X2, … Xk : bağımsız değişkenler
- Basit doğrusal regresyon denklemi: Y= bX+ a
- Çoklu regresyon denklemi: $Y = b_1X_1 + b_2X_2 + \cdots + b_kX_k + a$
- Çoklu korelasyon katsayısı (R)= Bir değişken ile (Y) bir grup değişken (X1, X2, … , Xk) arasındaki korelasyon katsayısıdır.

Örnek: Bir araştırmacı öğrencilerin duyuşsal özelliklerinin matematik başarılarını tahmin etmede (yordamada) etkili olup olmadığını incelemek istemektedir. Bu amaçla öğrencilere beş farklı ölçek uygulamıştır:

- Matematik özyeterliliği
- Matematiğe yönelik ilgi
- Başarısızlık atfetme
- Matematik çalışma etiği
- Matematik dersindeki davranışları

Araştırma sorumuz:

Matematik özyeterliliği, Matematiğe yönelik ilgi, Başarısızlık atfetme, Matematik çalışma etiği, Matematik dersindeki davranışları değişkenleri matematik başarısını yordar mı?

Regresyon denklemimiz:

Başarı=b1*Ozyet + b2*ilgi+ b3*b_atif + b4*etik + b5*davranis + a

Çoklu regresyon yöntemleri:

- Çoklu regresyon modellerinde modele eklenen değişkenler sonucu etkiler. Y değişkenini yordamak için kullanacağımız iki bağımsız değişkenimiz X1 ve X2 olsun. Modeli yalnızca X1 ile kurduğumuzda bir regresyon katsayısı elde ederiz. Modele X2 değişkenini eklediğimizde X1 değişkeninin katsayısı ve/veya manidarlığı değişebilir.
- Araştırmacının bir çok bağımsız değişkeni olduğunda ve karmaşık bir model kurması gerektiğinde bağımsız değişkenlerden bir veya birkaçını seçmesi gerekebilir.
- Modele eklenecek değişkenlere genellikle literatüre veya teoriye dayalı olarak karar verilir.
- Elimizde teoriden gelen bir bilgi olmadığında ise istatistiksel yöntemlerle de karar verilebilir.

Yöntemler:

- Hiyerarşik: Hangi değişkenlerin modele hangi sırayla eklenmesi gerektiğine araştırmacının (literatüre ya da önceki araştırmalarına göre) belirlediği analiz türüdür.
- Forced entery: Tüm değişkenlerin modele eklendiği yöntem (SPSS de enter olarak varsayılan yöntem)
- Stepwise: Modele hangi değişkenleri ekleyeceğimize ilişkin teorik bir bilgimiz olmadığında kullanırız. Bu model değişkenlerin hangi sırayla modele gireceğine matematiksel bir kritere bakarak karar verilir.
    - Backward: Tüm değişkenler modele eklenir. Etkili olmayan değişkenler sıraylar çıkarılır
    - Forward: Değişkenler sırasıyla modele eklenir. Eklenen değişkenlerin bağımlı değişken üzerinde anlamlılığına göre son modele karar verilir

REGRESYON MODELİMİZ NE KADAR DOĞRU?

- Model veriye ne kadar uyumlu?
- Model uç değerlerden etkileniyor mu?
- Model başka örneklemlere de genellenebilir mi?
- Uç değerler
- Artıklar
- Etkili değerler

Artıklar (Residuals)

- Verideki gerçek değerler ile kestirilen değerler arasındaki farklardır Y-$\hat{Y}$
- Modelin hatasıdır.
- Model iyi uyum gösterirse hatalar küçük olur.
- Artıkların yüksek olması veride uç değer olduğunun bir göstergesi olabilir.
- Artık değerler standartlaştırılmış veya standartlaştırılmamış olmak üzere iki şekilde raporlanabilir.
- Standartlaştırılmamış artıklar bağımlı değişkenle aynı birime sahiptir ancak hatanın büyüklüğünün yorumlanması zordur.

- Artık değerlerin standartlaştırılması iki şekilde yapılabilir:
  - standartlaştırılmış artık değerler: Gerçek değer ile tahmini değerlerin farkı onların standart sapma tahmini değerine bölünerek elde edilir.
  - studentleştirilmiş (studentized) artık değer: her bir nokta için ayrı ayrı standart sapma tahmini elde edilir ve farklı bu değere bölünerek elde edilir.
- Standartlaştırılmış Artık Değerlerin dağılımının ortalaması 0 standart sapması 1 dir. Problemli veriler veya uç değerler standartaştırılmış artıklara bakılarak elde edilebilir. (z puanları gibi)
- +-3.29 sınırı dışında kalan değerler genelde büyük artık değerleri olarak yorumlanır ve model uyumunu çok etkileyeceği söylenir. (farklı kaynaklarda farklı sınır değerleri görebilirsiniz.)

Etkili Değerler (Infuential)

- Veriden silindiğinde regresyon katsayılarını çok fazla etkileyen/değiştiren noktalara (kişilere) etkili değerler denir.
- Bu etkili değerleri belirlemek için farklı istatistikler yer almaktadır.
- DFFit: Orijinal tahmin değeri ile düzeltilmiş tahmin değeri arasındaki farktır.
- Mahanolobis uzaklığı: Her bir veri noktasının tahmin edilen değişkenlerin ortalamasından uzaklığını gösterir. Yaygın olarak kullanılır. Kritik değerleri örneklem büyüklüğü ve bağımsız değişken sayısına göre belirlenir.

  Serbestlik derecesi = bağımsız değişken sayısı -1 iken alfa 0.001 düzeyinde ki kare tablo değeri kritik değerlerdir. Bu değerden büyük olan uzaklıklar etkili değerdir. Analizden çıkarılmalıdır.

Varsayımlar

- Eşvaryanslılık: Her bir bağımsız değişken düzeyinde hataların varyansı sabit olmalı. Grafiklerde ZRESID olanı y-aksisine *ZPRED olanı da x-aksisine ekleyerek elde edeceğimiz grafik rastgele hatalar ve eş-varyanslılık varsayımlarını kontrol etmemize yardımcı olacaktır.
- Bağımsızlık (Independence): Bağımlı değişkenin her bir değerinin birbirinden bağımsız olduğu varsayılır.
- Doğrusallık (linearity): İlişkinin doğrusal olduğu varsayılır
- Çoklu bağlantı
- Hataların bağımsızlığı
- Hataların normalliği

Çoklu bağlantı (Multicolinearity)

- Bağımsız değişkenler arasında güçlü ilişkilerin olması durumuna çoklu bağlantı denir.
- Çoklu bağlantı değişkenler arasındaki ilişkilerin yüksek olması durumunda ortaya çıkar.
- Bu tür durumlarda söz konusu değişkenlerden bir yada birkaçının modelden çıkarılması önerilir.

Çoklu bağlantı göstergeleri

- Basit ikili korelasyon .80 ve üzeri olduğunda
- Çoklu regresyonda R^2 de değişim olmadığında
- İki değişken arsında korelasyon manidarken kısmi korelasyon katsayısının manidar olmaması
- Varyans artış faktörü,VIF>10 (1/ (1-R^2))
- Tolerans Değeri,TV<.10 (1-R^2)
- Koşul sayısı, CI > 30 (Bağımsız değişkenlerin ortak varyanslarının özdeğerlerine oranı)

Olduğu durumlarda çoklu bağlantı vardır. Değişkenlerden yüksek korelasyon gösterenlerden birini modelden çıkarmak gerekir.

Hataların bağımsızlığı

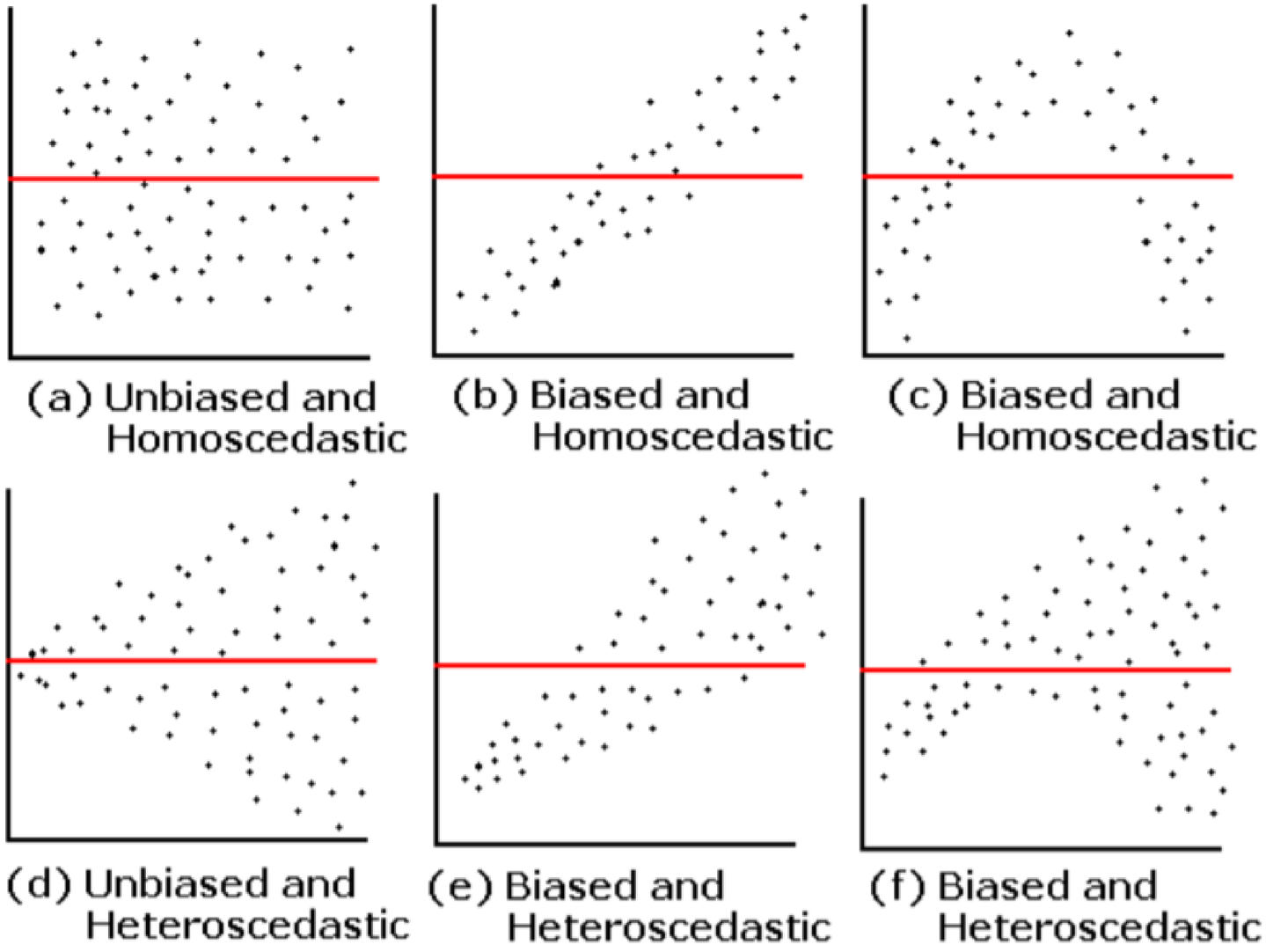
- Herhangi iki veri noktası için artık değerlerin bağımlı/ilişkili (korelasyonlu) olmaması gerekir.
- 0 ile 4 arasında değişen Durbin–Watson testi ile test edilebilir.
- 2 değeri korelasyonsuz olma durumunu gösterirken 2 den büyük ve küçük değerler negatif ya da pozitif korelasyonu gösterir.

Hataların Normalliği

- Modeldeki artık değerlerin rastgele ve normal dağılım gösterdiği varsayılır.
- Grafiksel yöntemlerle incelenebilir.
- *SRESID (y-axis) ile *ZPRED (x-axis) arasında oluşturulan grafik de eş-varyanslılık ihlalini göstermek için kullanılabilir.

Eşvaryanslılık grafikleri ve yorumlar:

Zpred vs zresid



Kategorik kestirici değişkenler

- Çoklu regresyonda bağımsız değişkenler genellikle sürekli değişkenlerdir.
- Kategorik olan değişkenlerin bağımlı değişken üzerindeki etkilerini incelemek istediğimizde bu değişkenler de regresyon modeline eklenebilir.
- İki kategorili bir bağımsız değişken 0 ve 1 şeklinde kodlanarak doğrudan regresyon modeline eklenebilir.
- Kategori sayısı ikiden fazla olduğunda ise kukla değişken (dummy variable) oluşturulur.

Kukla değişkenler:

- Kukla değişken sayısı kategori sayısı-1 kadardır. (k-1)
- Kukla değişken oluşturulurken:
  - k-1 tane kukla değişken oluşturun
  - Kategorilerden biri temel kategori olarak seçilir. (Diğer grupları karşılaştıracağınız gruptur. Eğer belirli bir hipoteziniz veya bir gruba ilişkin öngörünüz yok ise en kalabalık olan grubu seçin)
  - Temel aldığınız grubun değerini tüm kukla değişkenlerde (k-1 tane) 0 olarak belirleyin.
  - Birinci kukla değişkeniniz için birinci grubu 1 olarak diğer grupları 0 olarak kodlayın
  - İkinci kukla değişkeniniz için ikinci grubu 1 diğer grupları 0 olarak kodlayın
  - k-1 değişken için bunu tekrarlayın

Örn: Verimizde öğrencilere hangi alanda kariyer yapmak istediklerini soran bir KariyerPlan değişkenimiz var. Bu değişken üç kategorili: Matematik, Fen ve Sosyal. Bu değişkene ilişkin iki kukla değişkenimiz olacak. Temel kategoriyi matematik olarak seçersek aşağıdaki gibi bir kodlama yapmamız gerekir.

Kategoriler	Kukla Değişken 1	Kukla Değişken 2
Matematik	0	0
Fen	1	0
Sosyal	0	1

Lojistik Regresyon

- Bağımlı/ kestirilen değişkenin ikili puanlanan kategorik bir değişken olduğu bir çoklu regresyon modelidir.
- Y bağımlı değişken: iki kategorili
- X bağımsız değişkenler: sürekli veya iki kategorili

Örn:

- Öğrencilerin bir sınavdan geçtiğini veya kaldığını,
- Bir tümörün kanserli olup olmadığı,
- Bir müşterinin verilen krediyi ödeyip ödemeyeceği …

- Lojistik regresyonda kestirmeye çalıştığımız Y değişkeninin değerleri değil kategorilerine ilişkin olasılık değerleridir.
- Basit doğrusal regresyon denklemi: Y= bX+ a
- Çoklu regresyon denklemi: $Y = b_1X_1 + b_2X_2 + \cdots + b_kX_k + a$
- Lojistik regresyon denklemi ise:

$$P(Y) = \frac{1}{1 + e^{-(b_1X_1 + b_2X_2 + \cdots + b_kX_k + a)}}$$

- Çoklu regresyonun temel varsayımlarından birisi bağımlı değişken ile bağımsız değişkenler arasından doğrusal bir ilişki olmasıdır.
- Bağımlı değişken iki kategorili olduğunda bu ilişki doğrusal olmaz.
- Lojistik regresyon çoklu regresyon denklemine logaritmik bir dönüşüm yaparak doğrusallık varsayımından kurtarır.

- Denklem bu haliyle bize Y nin olma olasılığını (Belirli bir kategorinin olma olasılığını) verir.
- Olasılık 0 ile 1 aralığında değerler alır.
- Arka planda bu olasılıklara dayalı olarak regresyon denklemi ile bireyler tekrar sınıflandırılır.
- Analiz sonucunda:
  - modelin anlamlılığna ilişkin bir ki kare testi sonucu
  - doğru sınıflama yüzdesi
  - Her bir bağımsız değişkenin modelde varlığının anlamlı olup olmadığına ilişkin Wald istatistiği

Örnek: Bir araştırmacı öğrencilerin duyuşsal özelliklerinin matematik dersinden geçip geçememe durumlarını tahmin etmede (yordamada) etkili olup olmadığını incelemek istemektedir.

Bağımsız değişkenler: Matematik özyeterliliği,Matematiğe yönelik ilgi,Başarısızlık atfetme, Matematik çalışma etiği,Matematik dersindeki davranışları

Bağımlı değişken: Dersten geçme durumu