Çoklu Doğrusal Bağıntı (ÇDB)

* ÇDB, genel olarak açıklayıcı (bağımsız) değişkenler arasındaki lineer ilişkiyi ifade eder. Buradaki 'lineer' (doğrusal) kelimesi önemlidir.
* ÇDB bu ilişkinin lineer olmasıyla ilgilenir. Eğer bağımsız değişkenler arasındaki ilişki non-lineer bir form alıyorsa ÇDB'ye sebep olmaz.
* ÇDB sadece çoklu regresyonda ortaya çıkan bir problemdir. Bu nedenle verilen örnekler sadece çoklu doğrusal regresyon üzerinden verilecektir. ÇDB neden basit regresyonda gözlenmez?

Çoklu Doğrusal Regresyon

Hatırlatma

Model → $y = B_0 + B_1 X_1 + B_2 X_2 + \dots + B_K X_K + u$ (1)

Model indeksle yazıldı → $y_i = B_0 + B_1 X_{i1} + B_2 X_{i2} + \dots + B_K X_{iK} + u_i$ (2)

$y \Rightarrow$ bağımlı değişken

$X = \{X_1, X_2, \dots X_K\} \Rightarrow$ bağımsız değişkenler

$u \Rightarrow$ stokhastik hata terimi

$n =$ gözlem sayısı

$i \Rightarrow i.$ gözlem (indeks)

$B_0 \Rightarrow$ sabit terim

$B_J = \{B_1, B_2, \dots, B_K\} \Rightarrow$ kısmi eğim parametreleri

$J =$ kısmi eğim parametreleri için indeks

ÇDR4: Örneklemde (ve bu nedenle anakütlede) bağımsız değişkenlerin hiçbiri sabit değildir ve bağımsız değişkenler arasında tam doğrusal bağıntı yoktur.

↓

= Tam Çoklu Doğrusal bağıntı

↳ Herhangi bir X diğer x'lerin lineer bir kombinasyonu olarak yazılamaz.

* Bu varsayım x'lerin birbirleriyle ilişkili olmasına izin verir. İzin verilmeyen tam lineer bir ilişkinin olmasıdır.

↳ Ne demek?

* X-ler tam lineer ilişkili olursa SEKK parametre tahmincilerinin tahmini matematiksel olarak mümkün olmaz. (Daha sonra gösterilecek).

* Bu varsayıma göre bağımsız değişkenler ilişkili olabilir (non-lineer ve lineer olarak). X'ler arasında ilişkiye izin vermezsek çoklu regresyondan istediğimiz faydayı alamayız. Bu noktaya da daha sonra bir örnek ile değineceğiz.

* Daha önce 1 numaralı denklemle verilen modeli SEKK yöntemi ile tahmin ettiğimizi düşünelim

ÖRF $\hat{y}_i = \hat{B}_0 + \hat{B}_1 X_{i1} + \hat{B}_2 X_{i2} + \dots + \hat{B}_k X_{ik}$ (3)

* 3 numaralı denklemdeki parametre tahmincileri ve varyansları her hangi bir kısmi eğim parametresi $\hat{B}_J$ için şu şekilde yazılabilir.

$$\hat{B}_J = \frac{\sum_{i=1}^{n} \hat{r}_{iJ} y_i}{\sum_{i=1}^{n} \hat{r}_{iJ}^2} \quad (4) \qquad Var(\hat{B}_J) = \frac{\sigma^2}{SST_J(1-R_J^2)} \quad (5)$$

* Bu noktada anlaşılırlığı arttırmak için 1 numaralı denklemle verilen modeldeki bağımsız değişken sayısını 2-ye düşürelim ve ÖRF-yi (3. numaralı denklem ile verildi) yeniden yazalım. Ayrıca $\hat{B}_J$ yerine $\hat{B}_1$, 4. ve 5. denklemler yerine de 8. ve 9. denklemleri yazalım.

model $\Rightarrow y_i = B_0 + B_1 X_{i1} + B_2 X_{i2} + u_i$ (6)

ÖRF $\Rightarrow \hat{y}_i = \hat{B}_0 + \hat{B}_1 X_{i1} + \hat{B}_2 X_{i2}$ (7)

$$\hat{B}_1 = \frac{\sum_{i=1}^{n} \hat{r}_{i1} y_i}{\sum_{i=1}^{n} \hat{r}_{i1}^2} \quad (8) \qquad Var(\hat{B}_1) = \frac{\sigma^2}{SST_1(1-R_1^2)} \quad (9)$$

* Burada $\hat{r}_{i1}$, x_1-in x_2 üzerine regresyonundan elde edilen artıklardır.

$X_{i1} = \hat{\delta}_0 + \hat{\delta}_1 X_{i2} + \hat{r}_{i1}$ (10) $\longrightarrow$ $R_1^2 \Rightarrow$ bu modelin belirlilik katsayısı

$\longrightarrow$ $SST_1 \Rightarrow$ toplam kareler toplamı

$y_i = \hat{B}_0 + \hat{B}_1 \hat{r}_{i1}$ + kalıntı

$$SST_1 = \sum_{i=1}^{n} (X_{i1} - \bar{X}_1)^2$$

★ 6. numaralı denklem ile verilen modelde ÇDB4 varsayımı

• Basitçe bağımsız değişkenlerden hiçbiri kalan bağımsız değişkenlerin kombinasyonu olarak yazılamaz der.

• Resmi olarak ise aşağıdaki denklemi sağlayan, (0,0) haricinde hiçbir (λ_1, λ_2) ikilisi yazılamaz der.

$$\lambda_1 X_1 + \lambda_2 X_2 = 0$$

• Eğer (λ_1, λ_2) ikilisi (0,0) haricinde değerler alabilirse yani X_1 ve X_2 arasında lineer bağıntı varsa tam çoklu doğrusal bağıntı (ÇDB) problemi vardır ve 7. denklem tahmin edilemez (parametre tahmincileri ve varyansları hesaplanamaz). Aksi durumda X_1 ve X_2 bağımsızdır ve 7. denklem hesaplanabilir.

ÖR 1: $X_1 - 2X_2 = 0 \Rightarrow X_1 = 2X_2 \Rightarrow \begin{matrix}\lambda_1 = 1 \\ \lambda_2 = 2\end{matrix}$ ise

• Bu durumda tam ÇDB problemi var. Bağımsız değişkenler lineer olarak birbirine bağımlı. Hesap yapılamaz.

• Bu durumu önce mantık ve modelin kendisini kullanarak anlamaya çalışalım.

★ $y_i = B_0 + B_1 X_{i1} + B_2 X_{i2} + u_i$ — $X_1 = 2X_2$ yerine koy

$y_i = B_0 + B_1 2 X_{i2} + B_2 X_{i2} + u_i$

$y_i = B_0 + (2B_1 + B_2) X_{i2} + u_i$

$y_i = B_0 + \alpha X_{2i} + u_i$

$\boxed{\alpha = 2B_1 + B_2} \Rightarrow$ SEKK yöntemi ile α bulunabilir fakat B_1 ve B_2 değerleri tekil olarak bulunamazlar.

★ y = maaş, X_1 = eğitim süresi, X_2 = Eğitim süresinin 2 katını göstersin.

↳ Eğitim süresi ⇒ maaşı ne kadar etkiler?

$$y = B_0 + B_1 X_{1} + B_2 X_{2} + u$$

Aynı bilgi

Diğer herşey sabit iken

B_1'in yorumlamasını hatırlayın

"Ceteris paribus X_1 1 birim artınca y'nin ne kadar arttığını gösterir" ⇒ X_1 1 birim artınca X_2 nasıl sabit kalabilir?

* Hesaplamanın yapılamadığını göstermenin bir diğer yolu ise $\hat{B}_1$ ve $Var(\hat{B}_1)$-in formulünü kullanmaktır.

$$\hat{B}_1 = \frac{\sum \hat{r}_{i1} y_i}{\sum \hat{r}_{i1}^2} \quad (8) \qquad Var(\hat{B}_1) = \frac{\sigma^2}{SST_1 (1-R_1^2)} \quad (9)$$

$$X_{i1} = \hat{\delta}_0 + \hat{\delta}_1 X_{i2} + \hat{r}_{i1} \quad (10)$$

- Eğer X_1 ve X_2 arasında tam bir bağıntı varsa $\hat{\delta}_1 = 2$ bulunacak ve $\hat{r}_{i1}$'nın tüm değerleri sıfır çıkacaktır (10. denklem). Bu durumda da $R_1^2 = 1$ olacaktır.
- $\sum \hat{r}_{i1}^2 = 0$ olacak ve $\hat{B}_1$ tanımsız olacaktır.
- $R_1^2 = 1$ oldugundan $Var(\hat{B}_1)$ da tanımsız olacaktır.

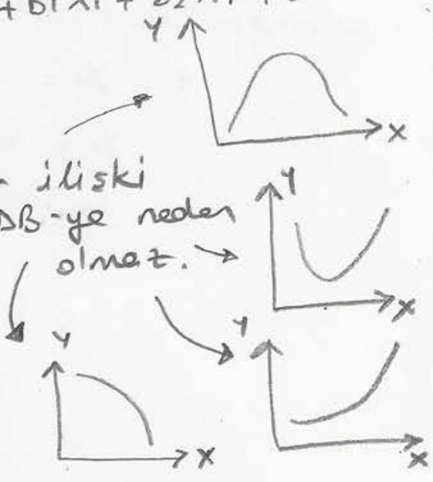
---

* Tam çoklu bağıntının ortaya çıktığı bazı örneklere bakalım. Ama bunun için 6. denklemi biraz değiştirelim

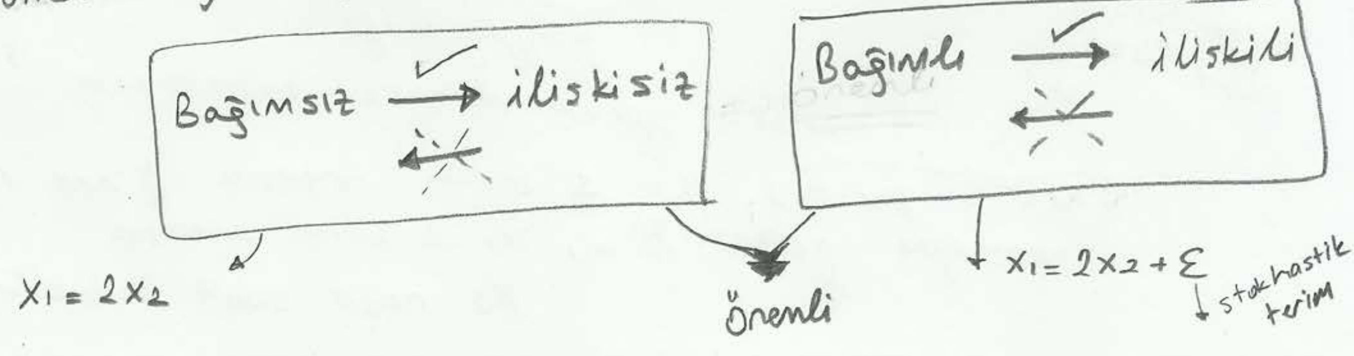
$$y_i = B_0 + B_1 X_{i1} + B_2 X_{i2} + B_3 X_{i3} + u_i$$

- $X_1 = 3X_2 \Rightarrow$ TÇDB var
- $X_1 = X_2 + X_3 \Rightarrow$ " "
- $X_3 = 2X_2 + 4 + 3X_1 \Rightarrow$ " "
- $X_1 = X_2^2 \Rightarrow$ TÇDB YOK $\Rightarrow$ Non-lineer ilişki TÇDB-ye neden olmaz.
- $\log(X_1) = \log(X_2^2) \Rightarrow$ TÇDB VAR
  $= 2\log(X_2)$
- $\log(X_1) = [\log(X_2)]^2 \Rightarrow$ TÇDB YOK

$y = B_0 + B_1 X_1 + B_2 X_1^2 + U$



* Not: TÇDB varsa değişkenler bağıntılıdır (yani bağımlıdır). Eğer değişkenler birbirine bağlı ise ilişkilidir fakat bunun tersi her zaman doğru değildir.



* ÇDB 4 varsayımı bağımsız değişkenler arasındaki bağımlılığı yani 1-e 1 bir ilişkiyi istemez, Fakat tamdan daha az bir ilişkiye izin verir.

• Zaten eğer bağımsız değişkenler arasında bir ilişkiye izin verilmezse (ya da yoksa) çoklu doğrusal regresyonun bir anlamı kalmaz ve tüm bağımsız değişkenler ayrı ayrı bağımlı değişken üzerine regres edilip parametreler hesaplanabilir.

ÖR 2: $Puan = B_0 + B_1 Harcama + B_2 Gelir + u$

Puan = öğrenci başarı puanı
Harcama = öğrenci ailesinin harcaması
Gelir = " " gelir

Bu regresyonda öğrencinin puanına ailesinin yaptığı harcama ve geliri nasıl etki eder bulmaya çalışıyoruz. Modele harcamayı koymamızın amacı: harcama ve gelirin ilişkili olduğunu biliyoruz ve bu nedenle geliri sabit tutarak harcamanın puan üzerindeki yalın etkisini hesaplamaya çalışıyoruz.

* Özetlemek gerekirse SEKK yönteminin uygulanabilmesi için tek şart tam çoklu doğrusal bağıntı olmamasıdır. Fakat bağımsız değişkenler arasında lineer ilişki olabilir.

* ÇDB-nin otokorelasyon ve değişen varyans probleminden farkı anakütlenin değil örneklemin bir sorunu olmasıdır. Bu nedenle ÇDB-nin varlığından ziyade önemi (gücü) anlam kazanır. ÇDB-nin gücündeki artışın nelere sebebiyet verdiği ÇDB-nin sonuçları kısmında incelenecektir.

ÇDB-nin Sebepleri

① Veri toplama metodu: bağımsız değişkenlerin dar bir aralıktaki değerlerden toplanmış olması.

② Model ya da örneklemdeki kısıtlar:

$$\text{Elektrik Tüketimi} = B_0 + B_1 \text{Gelir} + B_2 \text{Ev Büyüklüğü} + u$$

otomatikman bağlılık

③ Fonksiyonel formun yanlış seçilmesi:

$Y = B_0 + B_1 X_1 + B_2 X_1^2 + u \Rightarrow$ özellikle X-ler küçük iken

④ Gereğinden fazla tanımlanan model

ÇDB-nin Sonuçları

* Daha önceden belirtildiği gibi ÇDR4 varsayımı sadece tam ÇDB-ye izin vermez. Yani bağımsız değişkenler arasında lineer ilişki olabilir fakat modele olan etkisini anlayabilmek için ÇDB-nin gücü ölçülmelidir.

* ÇDB durumunda SEKK parametre tahmincileri hala sapmasız, etkin, doğrusal ve tutarlıdır. Yani hala DESTE-dir.

Doğrusal En iyi Sapmasız Tahmin Edici

* Bu durum ÇDB-nin gücü ne kadar büyük olursa olsun geçerlidir (tam ÇDB hariç)

* Peki, ÇDB-yi biz neden problem olarak görürüz? Bu soruya cevap vermek için 6. denklemdeki modeli ve ona bağlı denklemleri kullanalım

$$\text{model} \Rightarrow y_i = B_0 + B_1 X_{i1} + B_2 X_{i2} + u_i \quad (6)$$

$$\text{ÖRF} \Rightarrow \hat{y}_i = \hat{B}_0 + \hat{B}_1 X_{i1} + \hat{B}_2 X_{i2} + u_i \quad (7)$$

$$\boxed{Var(\hat{B}_1) = \frac{\sigma^2}{SST_1(1-R_1^2)}} \quad (9)$$

$$X_{i1} = \hat{\delta}_0 + \hat{\delta}_1 X_{i2} + \hat{r}_{i2} \quad (10)$$

$\hookrightarrow R_1^2 \Rightarrow$ bu modelin belirlilik katsayısı

$$SST_1 = \sum_{i=1}^{n} (X_{i1} - \bar{X}_1)^2$$

$$\boxed{Corr(X_1, X_2)^2 = R_1^2}$$

* X_1 ve X_2-nin arasındaki lineer ilişkinin artması yani korelasyonun artması R_1^2-nin artmasına neden olur.

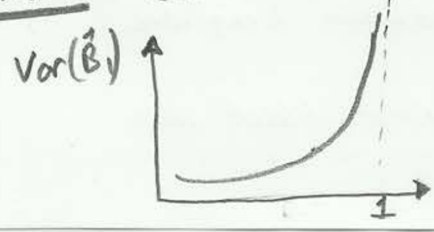
* $R_1^2 = 1$ tam ÇDB durumudur ve bu durumda hesap yapılamaz

* $R_1^2 = 0$ ise basit regresyon durumuna dönülür. Yani çoklu doğrusal regresyona gerek yoktur.

* R_1^2 değeri artmaya başladıkça $Var(\hat{B}_1)$ değeri artar ve R_1^2 yeteri kadar 1-e yaklaşınca sonsuz olur. Kısacası tahmin edilen $\hat{B}_1$ parametre tahmincisinin hassaslığı azalır.

Sonuç: ÇDB $\uparrow \Rightarrow R_1^2 \uparrow \Rightarrow Var(\hat{B}_1) \uparrow \Rightarrow$ hassaslık düştü $\rightarrow$ hassas bir tahmin yapmak zorlaşır.

$\hookrightarrow se(\hat{B}_1) \uparrow$



$\Rightarrow R_1^2$ arttıkça $Var(\hat{B}_1)$ parabolik olarak artar.

1-e hiç yaklaşmadığını varsayalım.

* $se(\hat{B}_1)\uparrow$ ise, bu durumda t-değerleri ve dolasıyla t-testleri etkilenir.

- B_1 için anlamlılık testi yaptığımızı düşünelim

$$H_0: B_1 = 0$$
$$H_1: B_1 \neq 0$$

$$t\text{-değeri} = \frac{\hat{B}_1 - B_1}{se(\hat{B}_1)} \Rightarrow t\text{-değeri} = \frac{\hat{B}_1}{se(\hat{B}_1)}$$

$se(\hat{B}_1)\uparrow \Rightarrow$ t-değeri $\downarrow \Rightarrow$ yanlış olmasına rağmen temel hipotezin yanlış olarak kabulune daha çok neden olur. Yani red etmeniz gereken Ho hipotezini red edemez ve B_1-i istatistiki olarak anlamsız bulursunuz.

tip II hata
yani B artar

* $se(\hat{B}_1)\uparrow$ ise, bu durumda parametre aralık tahmincisi de etkilenir. Aralık büyür.

$$\underbrace{\hat{B}_1 - t_{\alpha/2} \cdot \underbrace{se(\hat{B}_1)}_{\uparrow}}_{\downarrow} < B_1 < \underbrace{\hat{B}_1 + t_{\alpha/2} \cdot \underbrace{se(\hat{B}_1)}_{\uparrow}}_{\uparrow}$$

aralık genişledi

Aynı sonuç!

- Kısacası, GAB-nin gücü artarsa Ho hipotezi daha kolay kabul edilir hale gelir.

* R^2 olması gerekenden daha yüksek çıkar. Bağımsız değişkenler arasındaki yüksek lineer ilişki nedeniyle sağ tarafın sol tarafı etkileme gücü gerçek olmayan şekilde artar.

$$y_i = B_0 + B_1 X_{i1} + B_2 X_{i2} + u_i$$

X_1 ve X_2-nin Y üzerine olan etkisi artmış olarak gözlenir. Fakat bu yanlıştır.

* Yüksek ÇDB varlığında F-testi de etkilenir.

- modelin geneli için anlamlılık testi: F-testi

$H_0: B_1 = B_2 = 0$

H_1: H_0 doğru değil

$$F = \frac{R^2/K}{(1-R^2).(n-k-1)}$$

$R^2 \uparrow$ ise $\rightarrow$ $F \uparrow$

- sonuç: doğru olmamasına rağmen H_0 hipotezi yanlış olarak daha sık red edilir. Yani modelin geneli için istatistiki olarak anlamlı sonuç bulunur fakat model gerçekte anlamlı olmayabilir.

* F-testi ve t-testleri arasındaki tezat sonuca dikkat edilmeli.

$\rightarrow$ önemli $\Rightarrow$ Neden?

* Kurulan modeldeki SEKK parametre tahmincileri ve onların varyansları küçük veri değişimlerine karşı çok hassastır. Yani verilerden bir kaçının bile değişmesi sonucu değiştirebilir. Hatta parametre tahmincilerinin işareti beklentilerin tersine bile çıkabilir.

* Genel sonuç $\Rightarrow$ Yüksek ÇDB $\uparrow$ $\Rightarrow R_i^2 \uparrow \Rightarrow Var(\hat{B}_1) \uparrow \Rightarrow se(\hat{B}_1) \uparrow$

$\downarrow$

$t \downarrow, R^2 \uparrow, F \uparrow$

tezatlık