HAFTA 4

ÇOKLU DOĞRUSALLIK

Çoklu doğrusallık yoktur varsayımına;

1. Çoklu doğrusallığın niteliği nedir?
2. Çoklu doğrusallık gerçekten bir sorun mudur?
3. Uygulamada doğurduğu sonuçlar nelerdir?
4. Varlığı nasıl anlaşılır?
5. Çoklu doğrusallık sorununu hafifletmek için ne gibi düzeltici önlemler alınabilir? soruları ile cevap aranır.

Çoklu doğrusallık terimi önceleri bir regresyon modelinin bütün ya da bazı açıklayıcı değişkenleri arasında "tam" ya da kesin doğrusal ilişkinin varlığı anlamında idi. $X_1, X_2, \ldots, X_k$ bağımsız değişkenli regresyon modelinde k değişken arasında kesin bir doğrusal ilişkinin varlığı

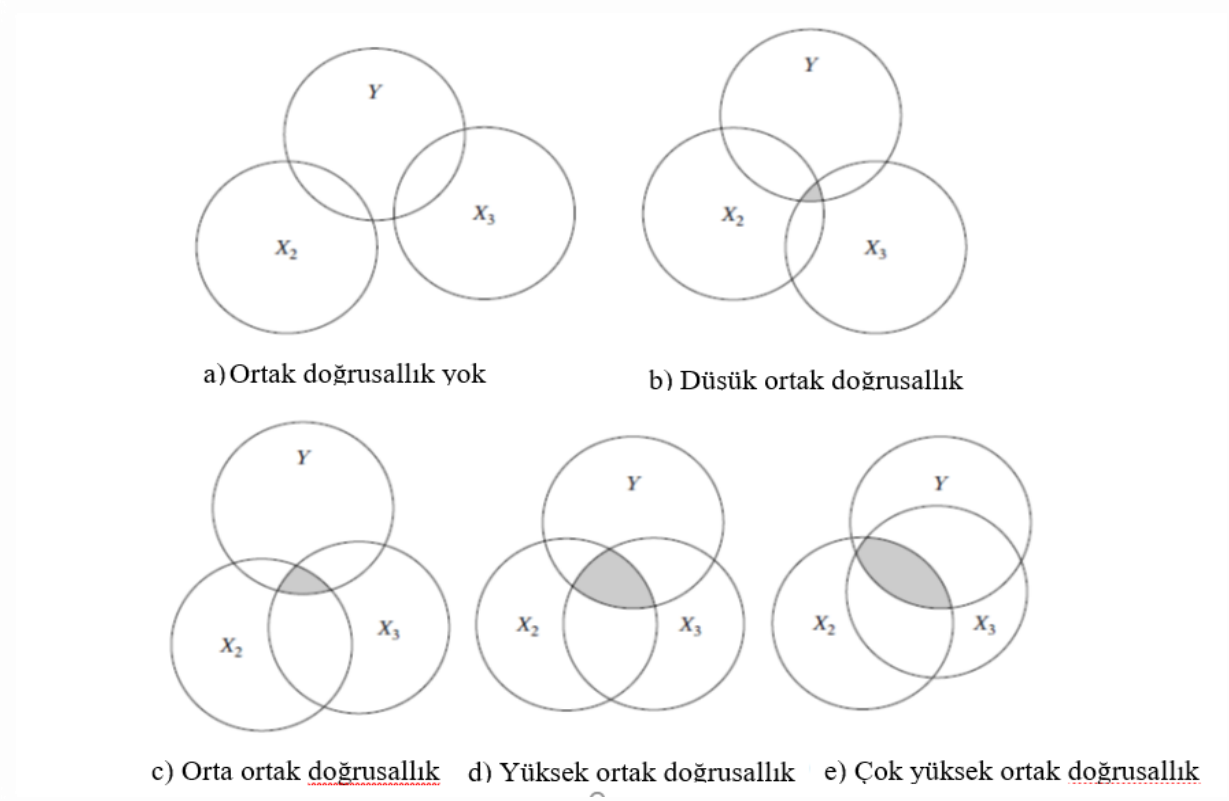
$$\lambda_1 X_1 + \lambda_2 X_2 + \ldots + \lambda_k X_k = 0 \quad \text{(tam)}$$

koşulunun sağlanması ile bulunabilir. Burada $\lambda_1, \lambda_2, \ldots, \lambda_k$ hepsi aynı anda sıfır olmayan sabitlerdir. Oysa bugün tam çoklu doğrusallığı ve X değişkenleri arasında tam olmasa da birbirleriyle ilişki içinde olduklarını gösteren koşul

$$\lambda_1 X_1 + \lambda_2 X_2 + \ldots + \lambda_k X_k + v = 0 \quad \text{(tamdan az)}$$

dır. Burada v olasılıklı hata terimidir.

- Çoklu doğrusallık tanımı X değişkenleri arasında sadece doğrusal ilişkilerdir. Doğrusal olmayan ilişkiler içerilmez.



Eğer çoklu doğrusallık tam ise X değişkenlerinin regresyon katsayıları belirsiz olup, bunların standart hataları sonsuzdur. Eğer çoklu doğrusallık tamdan az ise regresyon katsayıları belirlenebilmekle birlikte, (katsayılara oranla) büyük standart hatalar taşırlar. Bu da katsayıların büyük bir doğruluk ya da kesinlikle tahmin edilememeleri anlamına gelir.

Çoklu doğrusallığın bağlı olduğu etmenler:

1. Kullanılan veri derleme yöntemi: Sınırlı bir aralıkta örneklem alma
2. Modeldeki ya da örneklem alınan anakitledeki sınırlamalar
3. Model kurma
4. Aşırı belirlenmiş bir model: Modelin gözlem sayısından daha çok değişken içermesi.

– **Tam çoklu doğrusallık varken parametre tahmini:**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

regresyon modeli göz önüne alınıp, X_1 ve X_2 arasında çoklu doğrusallık olduğu durumda $X_2 = \lambda X_1$ (orijinden geçen regresyon doğrusu) ise

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon = \beta_0 + \beta_1 X_1 + \beta_2 \lambda X_1 + \varepsilon$$
$$= \beta_0 + \underbrace{(\beta_1 + \lambda\beta_2)}_{\beta_1^*} X_1 + \varepsilon = \beta_0 + \beta_1^* X_1 + \varepsilon$$

Burada $\beta_1^* = \beta_1 + \lambda\beta_2$ bilinen en küçük kareler yöntemi ile $\hat{\beta}_1^* = \hat{\beta}_1 + \lambda\hat{\beta}_2 = \dfrac{\sum_{i=1}^{n} x_{1i} y_i}{\sum_{i=1}^{n} x_{1i}^2}$ bulunur.

Görüleceği gibi iki bilinmeyenli tek denklem olmasından dolayı β_1 ve β_2 için tek çözüm bulunamamaktadır. Tam çoklu doğrusallık durumunda $\hat{\beta}_1$ ve $\hat{\beta}_2$ varyansları ile standart hataları ayrı ayrı sonsuzdur.

– **Tam olmayan ya da tama yakın çoklu doğrusallık varken parametre tahmini:**

Tam çoklu doğrusallık, uçlarda bir hastalık durumudur. X değişkenleri arasında genellikle tam bir doğrusallık ilişki yoktur, özellikle de iktisadi zaman serilerine ilişkin verilerde.

$$a_1 x_1 + a_2 x_2 + v = 0 \Rightarrow -a_2 x_2 = a_1 x_1 + v \Rightarrow x_2 = \underbrace{-\frac{a_1}{a_2}}_{\lambda} x_1 - \underbrace{\frac{1}{a_2}}_{v} v^* \Rightarrow x_2 = \lambda x_1 + v$$

burada $\lambda \neq 0$ ve v ise olasılıklı hata terimi olup, $\sum_{i=1}^{n} x_{1i} v_i = 0$ dır. Daha önce verdiğimiz şekillerden ilki dışındakiler tam olmayan ortak doğrusallığı gösterir. Bu durumda β_1 ve β_2 regresyon katsayıları tahmin edilebilir. β_1 parametresinin tahmini

$$\hat{\beta}_1 = \frac{\left(\sum_{i=1}^{n} y_i x_{1i}\right)\left(\sum_{i=1}^{n} x_{2i}^2\right) - \left(\sum_{i=1}^{n} y_i x_{2i}\right)\left(\sum_{i=1}^{n} x_{1i} x_{2i}\right)}{\left(\sum_{i=1}^{n} x_{1i}^2\right)\left(\sum_{i=1}^{n} x_{2i}^2\right) - \left(\sum_{i=1}^{n} x_{1i} x_{2i}\right)^2}$$ bulunur. $x_2 = \lambda x_1 + v$ ve $\sum_{i=1}^{n} x_{1i} v_i = 0$ olmak üzere

$$\hat{\beta}_1 = \frac{\left(\sum_{i=1}^{n} y_i x_{1i}\right)\left(\lambda \sum_{i=1}^{n} x_{1i}^2 + \sum_{i=1}^{n} v_i^2\right) - \left(\lambda \sum_{i=1}^{n} y_i x_{1i} + \sum_{i=1}^{n} y_i v_i\right)\left(\lambda \sum_{i=1}^{n} x_{1i}^2\right)}{\left(\sum_{i=1}^{n} x_{1i}^2\right)\left(\lambda^2 \sum_{i=1}^{n} x_{1i}^2 + \sum_{i=1}^{n} v_i^2\right) - \left(\lambda \sum_{i=1}^{n} x_{1i}^2\right)^2}$$ olarak elde edilir.

Benzer bir ifade ile $\hat{\beta}_2$ bulunabilir. v_i yeterince küçük diyelim ki sıfıra çok yakın olması durumunda $\hat{\beta}_2$ hemen hemen tam ortak doğrusallık sergileyecektir.

$\beta_1 + \lambda\beta_2$ tahmine edilebilir bir fonksiyon ve $(\beta_1 + \lambda\beta_2) = \hat{\beta}_1 + \lambda\hat{\beta}_2$ en küçük kareler tahmin edicisi BLUE olacaktır.

- **Çoklu doğrusallığın doğurduğu kuramsal sonuçlar:**

Klasik modelin varsayımları sağlandığında regresyon katsayılarının en küçük kareler tahmin edicileri BLUE'dur. Çoklu doğrusallık, tama yakın çoklu doğrusallıktaki gibi çok yüksek olsa bile en küçük kareler tahmin edicileri BLUE özelliklerini korumayı sürdürürler. Öyle ise neden çoklu doğrusallık önem kazanıyor. Aslında çoklu doğrusallık hiçbir regresyon varsayımını çiğnemez. Sapmasız, tutarlı tahminler bulunur, bunların standart hataları da doğru hesaplanır. Çoklu doğrusallığın tek etkisi küçük standart sapmalı katsayı tahminleri bulmayı zorlaştırmasıdır. Gözlem sayısı tahmin edilecek katsayı sayısının üstündeyse çoklu doğrusallık ortaya çıkar.

- **Çoklu doğrusallığın uygulamada doğurduğu sonuçlar:**

1. EEK tahmin edicilerin BLUE olmalarına karşın varyansları ve ortak varyansları büyüktür, bu da kesin tahmini güçleştirir.
2. 1. Sonuç nedeniyle güven aralıkları çok geniş olma eğilimindedir, bu da "sıfır" (yani ana kütledeki gerçek katsayısının sıfır olduğu) yokluk önsavlarının kolayca red edilememesine yol açar.
3. 1. Sonuç nedeniyle, bir ya da daha çok katsayının *t* oranları $\left(t = \frac{\hat{\beta}}{S_{\hat{\beta}}}\right)$ istatistik bakımından anlamsız olur.
4. Bir ya da daha çok katsayının *t* oranları istatistik bakımından anlamsız olmasına karşın, bütünün uyum iyiliğinin ölçüsü R^2 çok yüksek olabilir.

5. EKK tahmin edicileriyle onların standart hataları, verilerdeki değişimlere karşı duyarlı olabilirler.

- **Çoklu doğrusallığın var olup olmadığını aramak:**

  - Çoklu doğrusallık bir nitelik sorunu değil, nicelik sorunudur.
  - Çoklu doğrusallık, olasılıklı olmadıkları varsayılan açıklayıcı değişkenlerin koşullarıyla ilgili olduğuna göre ana kitlenin değil, örneklemin bir özelliğidir.

KURALLAR:

1. Yüksek R^2 ama anlamlı pek az t oranı:

Tipik bir "klasik" çoklu doğrusallık belirtisidir. ANOVA tablosundaki test istatistiğinin büyük ve R^2 nin oldukça yüksek olmasına karşın, parametre testlerinde bağımsız değişkenlerin modele katkılarının anlamsız çıkması çoklu doğrusallık (multicollinearity) olduğunun göstergesidir. Eğer açıklayıcı değişkenlerin Y üzerindeki etkilerinin tamamı birbirinden ayır edilemeyecek durumdaysa, çoklu doğrusallık ancak o zaman zararlı sayılabilir.

2. Açıklayıcı değişkenler arasında çifter çifter yüksek korelasyon:

İki açıklayıcı değişken arasındaki basit ya da sıfırıncı dereceden korelasyon katsayısı yüksekse diyelim 0.80'i aşıyorsa, o zaman çoklu doğrusallık ciddi bir sorundur. Sıfırıncı dereceden yüksek korelasyonlar ortak doğrusallık izlenimini verseler de herhangi belli bir durumda çoklu doğrusallığın olması için korelasyonların yüksek olmasına gerek bulunmamasıdır. Teknik olarak sıfırıncı dereceden yüksek korelasyonlar, çoklu doğrusallık için yeterli ama gerekli olmayan bir koşuldur. Çünkü sıfırıncı dereceden ya da basit korelasyonlar düşük (diyelim 0.50'nin altında) olsa bile çoklu doğrusallık bulunabilir. Bunu görmek için

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i;\ i = 1, 2, \ldots, n$$

modeli göz önüne alınsın ve $X_{3i} = \lambda_1 X_{1i} + \lambda_2 X_{2i}$ olduğunu varsayalım. Burada λ_1 ve λ_2 ikisi birden sıfır olmayan sabitlerdir. X_3'ün X_1 ve X_2'nin doğrusal bir fonksiyonu olduğuna göre X_3'ün X_1 ve X_2'ye göre regresyonunda belirlilik katsayısı için $R^2_{3.12} = 1$ eşitliğini verir.

$$R^2_{3.12} = \frac{r^2_{31} + r^2_{32} - 2r_{31}r_{32}r_{12}}{1 - r^2_{12}} = 1$$

olduğuna göre $r_{13} = 0.5,\ r_{32} = 0.5,\ r_{12} = 0.5$ alınırsa, yani korelasyon katsayıları çok da yüksek olmayan değerler iken $R^2_{3.12} = 1$ eşitliğini görmek güç değildir. Öyleyse ikiden çok açıklayıcı değişken içeren modellerde basit ya da sıfırıncı dereceden korelasyon, çoklu doğrusallığın varlığını gösteren yanılmaz bir gösterge sayılmaz. Yalnız iki açıklayıcı değişken varsa, sıfırıncı dereceden korelasyon elbette yeterlidir.

3. Kısmi korelasyonların incelenmesi:

Sıfırıncı dereceden korelasyonlara güven sorunu nedeniyle Farrar ile Glauber kısmı korelasyon katsayılarına bakılmasını önerirler. Kısmi korelasyon katsayılarının incelenmesi yararlı olmakta birlikte bunların çoklu doğrusallık için yanılmaz bir gösterge olmaları kesin değildir. Çünkü hem R^2, hem de kısmi korelasyon katsayıları yeterince yüksek olabilir.

4. Yan Regresyonlar :

Çoklu doğrusallık, bir ya da daha çok açıklayıcı değişkenin, öteki açıklayıcı değişkenlerin tam ya da yaklaşık doğrusal bileşimi olmasından doğduğuna göre hangi X değişkeninin öteki X değişkenleriyle ilişkili olduğunu bulmanın yolu, her bir X_i'nin öteki X değişkenlerine göre regresyonu bulup buna karşılık gelen, R_i^2 diyeceğimiz R^2 değerini hesaplamaktadır. Bu regresyonlardan her birine Y'nin X'lere göre olan asıl regresyonunun yanı sıra hesaplandıklarından , yan regresyon denir. Daha sonra sonra F ile R^2 arasında kurulan ilişkiden yararlanırsak;

$$F_i = \frac{R^2_{X_i.X_1X_2\cdots X_{i-1}X_{i+1}\cdots X_k} \Big/ (k-2)}{\left(1-R^2_{X_i.X_1X_2\cdots X_{i-1}X_{i+1}\cdots X_k}\right) \Big/ (n-k+1)}, \quad i=1,2,\ldots,k$$

değişkeni $k-2$ ve $n-k+1$ serbestlik dereceli F dağılımına uyar. n örneklem büyüklüğünü, k sabit terimle birlikte açıklayıcı değişken sayısını, $R^2_{X_i.X_1X_2\cdots X_{i-1}X_{i+1}\cdots X_k}$, X_i değişkeninin kalan X değişkenlerine göre regresyonundan bulunan belirlilik katsayısını gösterir. Hipotez testinde test istatistiği F, tablo değeri F^*'dan büyükse X_i'nin öteki X 'lerle ortak doğrusal olmadığını gösterir. Doğrusal değillerse X_i modelde kalır. Bütün yan R^2 değerlerini biçimsel olarak test etmek yerine "Klein'in parmak hesabı" benimsenebilir. Buna göre bir yan regresyondan bulunan R_i^2 bütünün, yani Y'nin bütün açıklayıcı değişkenlere göre regresyonunun R^2'sinden büyükse, çoklu doğrusallık ancak o zaman can sıkan bir sorun olabilir.

5. Özdeğerler ve koşul endeksi:

SAS çoklu doğrusallığa tanı koymak için özdeğerleri ve koşul endeksini kullanır.

$$k \text{ koşul sayısı} = \frac{\text{En büyük özdeğer}}{\text{En küçük özdeğer}}$$

$$\text{Koşul endeksi} = KE = \sqrt{k}$$

Eğer; $100 \le k \le 1000$ ise çoklu doğrusallık orta ya da güçlü derecededir.
$k > 1000$ ise çoklu doğrusallık ciddidir.
Eğer; $10 \le KE \le 30$ ise çoklu doğrusallık orta ya da güçlü derecededir.
$KE > 1000$ ise çoklu doğrusallık ciddidir.

6. Hoşgörü ve varyans şişirme çarpanı:

k değişkenli modelde kısmi regresyon katsayısının varyansı

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^{n} x_{ji}^2} \underbrace{\left(\frac{1}{1-R_j^2} \right)}_{VŞÇ_j}$$

dir. Burada varyans şişirme çarpanı $VŞÇ = \frac{1}{1-R_j^2}$ olarak tanımlanır. R_j^2, X_j'nin diğer X'lere göre regresyonundaki R^2'dir. R_j^2 bire doğru atarken, yani X_j'nin öteki açıklayıcı değişkenlerle ortak doğrusallığı artarken $VŞÇ_j$'de artar ve limitte sonsuz olur.

$VŞÇ$ çoklu doğrusallığın bir göstergesi olarak kullanılabilir. $VŞÇ$ değeri ne kadar yüksek ise X_j'de o kadar "güçlük çıkarıcı" ya da ortak doğrusal olmaktadır. $VŞÇ > 10$ ise $R_j^2 > 0.90$ olduğu ortaya çıkar. Yani, ortak doğrusallık oldukça yüksektir. Bazen çoklu doğrusallığı ölçmek için hoşgörü ölçüsü kullanılır.

Hoşgörü: $HOŞ_j = 1 - R_j^2 = \frac{1}{VŞÇ_j}$

X_j'nin diğer değişkenlerle çoklu doğrusallığı yoksa $HOŞ_j = 1$, tam ilişkiliyse $HOŞ_j = 0$ olur. Yüksek bir $VŞÇ$ ile ölçülmüş yüksek bir çoklu doğrusallık, zorunlu olarak yüksek standart hatalar doğurabilir.

Vurgulanan nokta: Küçük örneklemin ve açıklayıcı değişkenlerdeki düşük değişkenliğin de en az çoklu doğrusallık kadar ciddi sorunlar yaratabileceğidir.