Hafta #6

Çoklu Doğrusal Bağıntının Belirlenmesi

* ÇDB-nin sonuçlarını açıklarken bahsettiğimiz R^2 büyümesi, t ve F testlerinin birbirlerinden farklı sonuçlar vermesi ÇDB-nin göstergesi olarak görülebilir.

* ÇDB-nin etkili olduğuna kesin bir karar verebilmek için bazı kriterlere bakılmalıdır. Burada bu kriterleri inceleyeceğiz.

① Varyans Büyütme Faktörü (VIF)

* VIF (Variance Inflation Factor) ile parametre tahmincilerinin ve varyanslarının ÇDB ile gerçek değerlerinden ne derece uzaklaştığı belirlenir.

* ÇDB-nin modeldeki hangi açıklayıcı (bağımsız) değişkenden kaynaklandığının bulunmasını sağlar. Kısacası ÇDB-nin modeldeki her bir bağımsız değişkenin varyansını ne kadar arttırdığını gösterir.

* Daha önceden tanımladığımız 3 ve 5 numaralı denklemleri hatırlayalım.

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

$$Var(\hat{\beta}_J) = \frac{\sigma^2}{SST_J(1-R_J^2)} \quad (5)$$

⇒ 5 numaralı denklemden görüldüğü üzere varyanstaki büyüme σ^2, SST_J ve R_J^2'ye bağlıdır.

* $\sigma^2 \Rightarrow$ regresyonun varyansı, $\hat{\sigma}^2$ ile tahmin edilir. Eğer yeteri kadar data toplanabilirse (yani $n\uparrow$), $\hat{\sigma}^2$ azalır ve varyans düşer. $\Rightarrow \hat{\sigma}^2 \downarrow \Rightarrow$ hata payının azalmasıdır aslında

* $n\uparrow \Rightarrow SST_J\uparrow \Rightarrow SST_J = \sum(x_{ij} - \bar{x}_J)^2$
  $\hookrightarrow Var(\hat{\beta}_J)\downarrow$

- $SST_J = 0 \Rightarrow$ ne zaman gerçekleşir?
  $\hookrightarrow$ bu durum hangi varsayım ile ihtimal dışı bırakılmıştır.

* $R_J^2 \uparrow \Rightarrow Var(\hat{\beta}_J) \uparrow$

↳ R_J^2 hangi durumlarda artar? ⇒ ÇDB-nin gücü artmaya başlarsa!

↳ $R_J^2 = 1$ ⇒ Tam ÇDB durumu ⇒ Bu durum ÇDR4 ile ihtimal dışı bırakılır. Çünkü $Var(\hat{\beta}_J)$ tanımsız olur.

↳ $R_J^2 = 0$ ⇒ ÇDB yok. Bu durumda aslında çoklu regresyona gerek yoktur.

$$Var(\hat{\beta}_J) = \frac{\sigma^2}{SST_J(1-R_J^2)} = \frac{\sigma^2}{SST_J}$$

→ $y_i = \beta_0 + \beta_J x_{iJ} + u_i$ ⇒ bu modelde tahmin edilen $\hat{\beta}_J$ parametre tahmincisinin varyansı.

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_J x_{iJ}$

* Kısacası $Var(\hat{\beta}_J)$-nin artmasına neden olan tek şey ÇDB değil veri sayısının yeteri kadar büyük olmamasıdır. ÇDB varlığında bile eğer yeteri kadar verimiz varsa ($n \uparrow$) $Var(\hat{\beta}_J)$ yeteri kadar küçük olabilir.

* Aslında ÇDB varlığından endişelenmek örneklem büyüklüğünün yeteri kadar büyük olmamasından ($n \downarrow$) endişelenmekle aynıdır. Bu nedenle ÇDB bazı ekonometriciler tarafından (Arthur Goldberger) micronumerosity (küçük örneklem problemi) olarak tanımlanır.

* Varyanstaki büyümenin ÇDB-den kaynaklanıp kaynaklanmadığını ve etkisinin önemli olup olmadığını belirlemek için VIF kriterinden faydalanılır.

$$VIF(\hat{\beta}_J) = \frac{1}{1-R_J^2}$$

⇒ incelenen modelde kaç tane bağımsız değişken varsa o kadar VIF hesaplanabilir.

yardımcı model x_J için
→ $x_J = \delta_0 + \delta_1 x_1 + \dots + \beta_{J-1} x_{J-1} + \beta_{J+1} x_{J+1} + \dots + \beta_k x_k + u$ → R_J^2 bu modelden hesaplanır.

ana model
→ $y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + u$

* Bazen ana model ve yardımcı modellerden elde edilen R^2-lerin karışmaması için aşağıdaki gösterim kullanılabilir.

ana model $\rightarrow$ $y = B_0 + B_1X_1 + B_2X_2 + \dots + B_kX_k + u \rightarrow R^2_{y,x_1,x_2,\dots,x_k}$

k tane bağımsız değişken var ve k tane VIF değeri hesaplanabilir.

$$VIF_1 = \frac{1}{1-R^2_{x_1,x_2,\dots x_k}} \; ; \; VIF_2 = \frac{1}{1-R^2_{x_2,x_1,\dots,x_k}} \; ; \dots ; \; VIF_k = \frac{1}{1-R^2_{x_k,x_1,x_2,\dots x_{k-1}}}$$

* VIF kriterinin hangi değeri aldığında ÇDB-nin etkili olacağı konusunda kesin bir kural yoktur. ÇDB-nin etkili olup olmadığı araştırmacı tarafından verilir.

$VIF_J > 5$ ya da $VIF_J > 10$ ise ÇDB etkilidir.

$$VIF = \frac{1}{1-R_J^2} > 5 \qquad VIF_J = \frac{1}{1-R_J^2} > 10$$

$$R_J^2 > 0{,}8 \qquad R_J^2 > 0{,}9$$

* VIF ne kadar büyük ise ÇDB o kadar ciddidir.
* $VIF = 1$ ise modelde ÇDB yoktur.

(2) Yardımcı Regresyon Modelleri için F testi

* VIF formülünde kullandığımız R_J^2 belirlilik katsayılarından yararlanılır.
* Sırası ile her bağımsız değişken diğerleri üzerine regres edilir. Oluşturulan bu modellere yardımcı regresyon denir.
* Oluşturulan bu modellerin belirlilik katsayıları kullanılarak F-istatistiği hesaplanır. Daha sonra bu yardımcı modelleri kullanarak modelin geneli için anlamlılık testi yapılır.
* F-testinde $\rightarrow$ H_0: bağımsız değişkenler arasında ilişki yok $\rightarrow$ ÇDB YOK

  H_1: bağımsız değişkenler arasında ilişki var $\rightarrow$ ÇDB VAR ve etkili ana model $\rightarrow$ $y = B_0 + B_1 X_1 + B_2 X_2 + \dots + B_K X_K + u$

K tane yardımcı regresyon
$$\left\{\begin{aligned} X_1 &= \delta_0 + \delta_2 X_2 + \delta_3 X_3 + \dots + \delta_K X_K + \varepsilon \\ X_2 &= \alpha_0 + \alpha_1 X_1 + \alpha_3 X_3 + \dots + \alpha_K X_K + e \\ &\vdots \\ X_K &= \theta_0 + \theta_1 X_1 + \theta X_2 + \dots + \theta_{K-1} X_{K-1} + v \end{aligned}\right\}$$
Her biri için R_J^2 hesapla

1. yardımcı regresyon için $\rightarrow$ $H_0: \delta_2 = \delta_3 = \dots = \delta_K$ $\rightarrow$ ÇDB YOK
H_1: H_0 doğru değil $\Rightarrow$ ÇDB VAR (X_1'den dolayı)

* Her yardımcı regresyon modeli için R_J^2 hesaplanır ve F-ist oluşturulur.

$$F\text{-ist}_J = \frac{R^2_{X_J, X_1, X_2, \dots X_K} \Big/ (k-1)}{\left(1 - R^2_{X_J, X_1, X_2, \dots, X_K}\right) \Big/ (n-k)}$$

k: ana modeldeki toplam bağımsız değişken sayısı

* Yardımcı regresyonlarda bir bağımsız değişken eksildiğinden bir parametre eksilmektedir. Bu nedenle serbestlik dereceleri de birer azalmaktadır.

* $F\text{-ist}_J$ değeri f-tablo (f-kritik) değeri karşılaştırılır.

F-kritik: $F_{\alpha, k-1, n-k}$

α = anlamlılık düzeyi
sd1 = k−1 $\Rightarrow$ paydaki serbestlik derecesi
sd2 = n−k $\Rightarrow$ paydadaki serbestlik derecesi

* Eğer $F\text{-ist}_J >$ F-kritik ise H_0 red edilir ve modeldeki X_J değişkeninden dolayı ÇDB olduğuna karar verilir.

* Çözüm ÇDB-ye neden olan bağımsız değişkenlerin modelden çıkarılması olabilir. Ama dikkat edilmelidir ki bu ihmal edilmiş değişken problemine neden olabilir. İhmal edilmiş değişken problemi ÇDB probleminden daha fazla soruna neden olur.

$\rightarrow$ Neden?

ÇDB $\rightarrow$ sapmasız, etkin $\rightarrow$ ama varyans şişmiş, tutarlı

ihmal edilmiş değişken $\rightarrow$ sapmalı, etkin $\rightarrow$ varyans küçük, tutarlı

* Bu testte ÇDB-ye neden olan bağımsız değişkenler bu değişkenlere ait yardımcı regresyon üzerine ayrı ayrı F testi yapılarak belirlenir.

③ Klein Kriteri

* Genellikle destekleyici bir kriter olarak kullanılır.
* Yardımcı regresyonların belirlilik katsayısı R_J^2 hesaplanarak ana modelin R^2-si ile karşılaştırılır.

$R_J^2 > R^2$ ise ÇDB ciddi boyuttadır ve parametreler güvenirliklerini kaybederler.

④ Theil-m Ölçüsü

* Bağımlı değişkenle bağımsız değişken arasındaki ilişkiye yani R^2-ye dayanan bir ölçüttür.
* önceki iki yöntem gibi ana regresyon ve yardımcı regresyonlar kullanılır. Fakat bu yöntemde ana modelden bir tane bağımsız değişken çıkartılarak sırası ile yardımcı regresyon oluşturulur.
* Theil-m ölçüsü her bağımsız değişken için hesaplanmayan genel bir ölçüttür. => VIF ve yardımcı regresyon modelleri için F testine göre bu yönüyle farklıdır.

Ana model => $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon \rightarrow R^2$

1. yardımcı model => $y = \delta_0 + \delta_2 X_2 + \dots + \delta_k X_k + e \rightarrow R^2_{-1}$

2. yardımcı model => $y = \alpha_0 + \alpha_1 X_1 + \alpha_3 X_3 + \dots + \alpha_k X_k + u \rightarrow R^2_{-2}$

⋮

k. yardımcı model => $y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \dots + \theta_{k-1} X_{k-1} + \varepsilon \rightarrow R^2_{-k}$

* Yardımcı regresyonlar için R^2_{-J} ve ana regresyon için R^2 değeri hesaplanıp m ölçüsü oluşturulur.

$$m = R^2 - \sum_{J=1}^{k} \left(R^2 - R^2_{-J}\right) \rightarrow (+) \text{ ve } (-) \text{ değerler alabilir.}$$

★ $\mu = 0$ ise ÇDB etkili değildir. Yardımcı regresyonların bireysel katkılarının toplamı R^2-ye eşitlenir

X_1 ve X_2 değişkenli için yazılırsa

$$\mu = R^2 - (R^2 - R^2_{-1}) - (R^2 - R^2_{-2})$$

$$\mu = R^2_{-1} + R^2_{-2} - R^2$$

★ $\mu = R^2 - \sum_{J=1}^{k} \underbrace{(R^2 - R^2_{-J})}_{}$

h_J olarak tanımlayalım

$m_L = R^2 - k.(\max h_J) \Rightarrow$ alt limit

$\mu_U = R^2 - k.(\min h_J) \Rightarrow$ üst limit $\longrightarrow$ k = bağımsız değişken sayısı

- $\mu_L < \mu < \mu_U$ ise ÇDB önemli değil
  aksi durumda ÇDB önemlidir

★ Gözlem sayısı 60-dan daha büyük ise VIF daha yararlıdır.

(5) Şartlı Sayı Kriteri

★ $X'X$ bağımsız değişkenler kullanılarak oluşturulan bu matrisin determinantı bize ÇDB hakkında bilgi verir.

$$Y = X\beta + \varepsilon$$

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} \quad X = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix}_{n\times(k+1)} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}_{(k+1)\times 1} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}$$

$$\hat{\beta} = \underbrace{(X'X)^{-1}}_{}X'Y$$

$\longrightarrow$ buradaki $X'X$ matrisinin tersinin alınabilmesi lazım. Bu matrisin tersinin alınabilmesi için tam çoklu doğrusal bağıntının modelde olmaması lazım.

* Şartlı sayı kriteri, bu $X'X$ matrisinin birim köklerinden (eigenvalue), yani özdeğerlerinden yararlanılarak oluşturulur.
* Bu kriter regresyonun bir bütün olarak ÇDB-den zarar görüp görmediğini ortaya çıkarır.
* Hangi bağımsız değişkenin ÇDB-ye neden olduğu hakkında fikir vermez.

$$\underset{\substack{\downarrow \\ \text{şartlı sayı} \\ \text{(conditional number)}}}{CN} = \sqrt{\frac{\max \lambda_i}{\min \lambda_i}} = \sqrt{\frac{\lambda_1}{\lambda_2}} = \sqrt{k}$$

$\lambda_1 = \max$ özdeğer

$\lambda_2 = \min$ özdeğer

* Hesaplanan CN değeri bazı belirli kriterler ile karşılaştırılarak ÇDB derecesine karar verilir.

$k < 10 \Rightarrow$ ÇDB ciddi değil

$10 < k < 30 \Rightarrow$ ÇDB orta derece ciddi

$30 < k \Rightarrow$ ÇDB ciddi

ÇDB-nin Düzeltilmesi

① Ön bilgilerin kullanılması: ÇDB-ye neden olan bağımsız değişkenlere ait bilgiler biliniyorsa bu bilgiler kullanılarak ÇDB düzeltilebilir. Teoriden ya da daha önceki çalışmalardan yararlanılır.

$$y = B_0 + B_1 X_1 + B_2 X_2 + \varepsilon$$

$B_2 = 0{,}3\, B_1$ olarak biliniyorsa

$$y = B_0 + B_1 X_1 + 0.3 B_1 X_2 + \varepsilon$$

$$y = B_0 + B_1 (X_1 + 0.3 X_2) + \varepsilon$$

$$y = B_0 + B_1 X^* + \varepsilon \longrightarrow X^* = X_1 + 0.3 X_2$$

* B_1 bulunduktan sonra B_2-de bulunabilir.

(2) Gözlem sayısını arttırmak → $\downarrow Var(\hat{B}_J) = \dfrac{\sigma^2}{SST_J(1-R_J^2)}$

$\hookrightarrow SST_J \uparrow$ ↗

(3) Yatay kesit ve zaman serisi verilerinin birleştirilmesi. Yani kısacası panel veri kullanımı.

Y = talep P = fiyat I = gelir

$\ln Y = B_0 + B_1 \ln P + B_2 \ln I + u$ → talep fonksiyonu zaman serisi ile

P ve I arasında yüksek korelasyon var

↳ eğer B_2 anket verilerinden yatay kesit ile tahmin edilir ve üstteki denklemin içine koyulursa.

$$\ln Y - B_2 \ln I = B_0 + B_1 \ln P + u$$

$$\underbrace{\ln Y - B_2 \ln I}_{y^*} = B_0 + B_1 \ln P + u$$

(4) Çoklu doğrusal bağıntıya neden olan değişkenleri modelden çıkarmak. Dikkat!!!

(5) Değişkenlerde dönüşümün (transformasyon) yapılması
↓
ilk farkların alınması
log alınması
büyüme alınması

(6) Modele yeni denklemler eklenmesi => eş anlı denklemler

$y = B_0 + B_1 X_1 + B_2 X_2 + u$ => enflasyon denklemi (modeli)

y ↓ TÜFE, X_1 ↓ Para Arzı, X_2 ↓ GSYH; Para Arzı ↔ GSYH ilişki

=> eş anlı çözüm.

?

$X_2 = \alpha_0 + \alpha_1 X_1 + \alpha Y + u_2$

ilişki

(7) Ridge regresyon yönteminin kullanılması: Gerekli olan tüm değişkenleri modele katar ve gereksiz değişkenlerin belli aşamalarla modelden atılmasına yarar