ÇOK DEĞİŞKENLİ REGRESYON ANALİZİ: TAHMİN

Hüseyin Taştan[1]

[1]Yıldız Teknik Üniversitesi
İktisat Bölümü

Ders Kitabı:
Introductory Econometrics: A Modern Approach (2nd ed.)
J. Wooldridge

17 Ekim 2012

Çok Değişkenli Regresyon Analizi (Multiple Regression Analysis)

- Basit regresyonda kilit varsayım olan SLR.3 varsayımı çoğu zaman gerçekçi olmayan bir varsayımdır. SLR.3: y'yi etkileyen tüm diğer faktörler x ile ilişkisizdir (ceteris paribus).
- Çoklu regresyon analizinde bağımlı değişkeni (y) eşanlı (simultaneously) olarak etkileyen pek çok etkeni kontrol edebiliriz. Zira, çok sayıda açıklayıcı değişken (x) kullanabileceğiz.
- Modele yeni değişkenler ekleyerek y'deki değişmenin daha büyük bir kısmını açıklayabiliriz. Yani, y'nin tahmini için daha üstün modeller geliştirebiliriz.
- Çoklu regresyonda regresyonun biçimini (functional form) belirlemede çok daha geniş olanaklara sahip olacağız.

Çoklu Regresyon Modeli Örnekler

İki Açıklayıcı Değişkenli Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

İki Açıklayıcı Değişkenli Ücret Denklemi

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

$wage$: saat başına ücretler (dolar); $educ$: eğitim düzeyi (yıl); $exper$: tecrübe düzeyi (yıl)

- Burada, β_1, ücretleri etkileyen diğer tüm faktörleri sabit tuttuğumuzda, eğitimin ücretlere etkisini ölçer.
- β_2 ise, benzer şekilde tecrübenin ücretlere ceteris paribus etkisini gösterecektir.
- Bu regresyonda tecrübeyi sabit (fixed) tutarak eğitimin ücretlere katkısını ölçebiliyoruz. Basit regresyonda bu olanak yoktu. Sadece *educ ile u ilişkisizdir* diye varsayıyorduk.

Çoklu Regresyon Modeli Örnekler

Sınav başarı notu ve aile geliri

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 avginc + u$$

$avgscore$: ortalama sınav sonucu; $expend$: öğrencinin eğitim harcamaları; $avgincome$: ortalama aile geliri

- Eğer aile gelirini (avginc) regresyona doğrudan sokmaz isek, onu, u'nun içinde ele almış olacağız.
- Aile geliri öğrencinin harcaması (expend) ile yakından ilişkili olduğundan, bu halde, x (harcama) ile u ilişkili olacak ve kilit varsayımımız, SLR.3, ihlal edilecekti. Bu ise β_1'in sapmalı (biased) tahmin edilmesine yol açacaktı.
- Avginc değişkenini modele ekleyerek onu doğrudan kontrol etme olanağına kavuştuk.

Çoklu Regresyon Analizi

- Çoklu regresyon, modelin fonksiyonel biçimini genelleştirmeye izin verir.
- Ailelerin tüketimini (consumption) gelirlerinin (income) karesel (quadratic) bir fonksiyonu olarak ifade edelim:

$$cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u$$

- Burada $x_1 = inc$ ve $x_2 = inc^2$
- Bu regresyonda β_1'in yorumu farklı olacaktır. Çünkü, gelirin karesini (inc^2) sabit tutarak gelirin tüketim üzerindeki etkisini ölçemeyiz. Gelir değişirse karesi de değişir.
- Burada gelirdeki bir birim değişmenin tüketim üzerindeki etkisi, yani *marjinal tüketim eğilimi* (marginal propensity to consume) şuna eşittir:

$$\frac{\Delta cons}{\Delta inc} \approx \beta_1 + 2\beta_2 inc$$

- Marjinal tüketimi gelir düzeyine bağlıdır.

Çoklu Regresyon Modeli Örnekler

İki Açıklayıcı Değişkenli Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

- İki açıklayıcı değişkenli modelde, "u'nun x'lerle ilişkisiz olması" varsayımını şu şekilde formüle edeceğiz:

$$\mathrm{E}(u|x_1, x_2) = 0$$

- Yani, x_1 ve x_2'nin kitledeki (population) tüm kombinasyonları için u'nun beklenen değeri sıfırdır.
- Örneğin iki değişkenli ücret denkleminde

$$E(u|educ, exper) = 0$$

- Bu ücretleri etkileyen diğer faktörlerin (u) ortalama olarak $educ$ ve $exper$ ile ilişkisiz olduğu anlamına gelir.
- Örneğin, doğuştan gelen yetenek (ability) u'nun bir parçası ise, ortalama yetenek düzeyi, çalışanlar kesiminde eğitim ve tecrübenin tüm kombinasyonlarında aynıdır (sabittir).

Çoklu Regresyon Modeli Örnekler

İki Açıklayıcı Değişkenli Model: u'nun x'lerle ilişkisiz olması

$$\mathrm{E}(u|x_1, x_2) = 0$$

- Test sonuçları ve ailenin geliri modelinde bu varsayım

$$E(u|expend, avginc) = 0$$

- Yani, test skorlarını etkileyen diğer faktörler (okula ya da öğrenciye özgü karakteristikler vs.), ortalama olarak, expend ve avginc değişkenleriyle ilişkisizdir.
- Karesel tüketim fonksiyonunda bu varsayım:

$$E(u|inc, inc^2) = E(u|inc) = 0$$

- Burada inc biliniyorken inc^2 otomatik olarak bilineceğinden ayrıca koşullu beklenti içinde yazmaya gerek yoktur.

k Açıklayıcı Değişkenli Regresyon Modeli

Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

- Bu modelde k açıklayıcı değişken, $k + 1$ bilinmeyen β parametresi vardır.
- Hata terimi, daha önce tanımlandığı gibi, x'ler dışında modele dahil edilmemiş tüm faktörlerin ortak etkisini temsil etmektedir.
- Modele ne kadar çok x değişkeni eklenirse eklensin dışarıda bırakılmış ya da gözlenemeyen faktörler her zaman olacaktır.

k Açıklayıcı Değişkenli Regresyon Modeli

- Herhangi bir parametre, β_j diyelim, diğer x'ler ve u'da içerilen faktörler sabitken ($\Delta u = 0$), x_j'deki bir birimlik değişmenin y'de yaratacağı değişmeyi gösterir.
- Ancak x'ler arasında doğrusal olmayan özellik varsa bu yorum değişir. Örneğin aşağıdaki modeli düşünelim:

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 ceoten + \beta_3 ceoten^2 + u$$

$ceoten$: yöneticinin aynı işyerinde çalışma süresi (yıl), kıdemi (tenure)

- Burada β_1 maaşların satış esnekliğidir. Diğer herşey sabitken (ceteris paribus), satışlarda meydana gelen %1 artışın yönetici maaşlarında yaratacağı yüzde değişmedir
- Ancak β_2 kıdemde bir yıl artış olduğunda maaşlarda ortaya çıkan yüzde değişimi göstermez. Karesel terimi de dikkate almak zorundayız.

k Açıklayıcı Değişkenli Regresyon Modeli

Sıfır Koşullu Ortalama Varsayımı

$$\mathrm{E}(u|x_1, x_2, \ldots, x_k) = 0$$

- Bu varsayım hata teriminin açıklayıcı değişkenlerle ilişkisiz olduğunu söylemektedir.
- Eğer u x'lerden biriyle ilişkiliyse OLS tahmin edicileri sapmalı (biased) olur. Bu durumda tahmin sonuçları güvenilir olmaz.
- İhmal edilmiş, yani dışarıda bırakılmış önemli bir değişken varsa, bu varsayım sağlanmayabilir. Bu da sapmaya yol açacaktır.
- Bu varsayım aynı zamanda fonksiyon kalıbının da doğru kurulduğu anlamına gelir.

k Açıklayıcı Değişkenli Regresyon Modeli: SEKK-OLS Tahmini

Örneklem Regresyon Fonksiyonu - SRF

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

Sıradan En Küçük Kareler (Ordinary Least Squares - OLS) tahmin edicileri kalıntı kareleri toplamını (SSR) en küçük yapar:

OLS amaç fonksiyonu

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik})^2$$

Birinci sıra koşullarından elde edilen $k+1$ denklemin çözümünden OLS tahmin edicileri $\hat{\beta}_j$'ler bulunur.

k Açıklayıcı Değişkenli Regresyon Modeli: SEKK-OLS Tahmini

OLS Birinci Sıra Koşulları

$$\begin{aligned}
\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik}) &= 0 \\
\sum_{i=1}^{n} x_{i1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik}) &= 0 \\
\sum_{i=1}^{n} x_{i2} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik}) &= 0 \\
&\vdots \quad \vdots \quad \vdots \\
\sum_{i=1}^{n} x_{ik} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik}) &= 0
\end{aligned}$$

k Açıklayıcı Değişkenli Regresyon Modeli: SEKK-OLS Tahmini

OLS ve Momentler Yöntemi

OLS birinci sıra koşulları aşağıdaki popülasyon moment koşullarının örnekleme karşılıkları olarak düşünülebilir:

$$\begin{aligned} \mathsf{E}(u) &= 0 \\ \mathsf{E}(x_1 u) &= 0 \\ \mathsf{E}(x_2 u) &= 0 \\ \vdots &\ \vdots \ \vdots \\ \mathsf{E}(x_k u) &= 0 \end{aligned}$$

Örneklem moment koşullarının (OLS birinci sıra koşulları) tek çözüm vermesi için gerekli varsayımları daha sonra inceleyeceğiz.

Regresyonun Yorumu

İki Açıklayıcı Değişkenli Durum

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

- Eğim parametre tahminleri, $\hat{\beta}_j$, açıklayıcı değişkenlerin y üzerindeki kısmi ya da ceteris paribus etkilerini verir.
- $\hat{\beta}_1$'nın yorumu: x_2 sabitken, yani $\Delta x_2 = 0$

$$\Delta \hat{y} = \hat{\beta}_1 \Delta x_1$$

- Benzer şekilde x_1 sabitken $\hat{\beta}_2$'nın yorumu

$$\Delta \hat{y} = \hat{\beta}_2 \Delta x_2$$

Örnek: Üniversite Başarısını Belirleyen Faktörler, gpa1.gdt

$colGPA$ Tahmin Sonuçları

$$\widehat{colGPA} = 1.29 + 0.453\ hsGPA + 0.0094\ ACT$$

$n = 141$ öğrenci, $colGPA$: üniversite genel not ortalaması (4 üzerinden puan), $hsGPA$: lise not ortalaması, ACT: genel yetenek sınav sonucu

- Sabit terim $\hat{\beta}_0 = 1.29$ olarak tahmin edilmiş. $hsGPA = 0$ ve $ACT = 0$ olduğunda modelce tahmin edilen üniversite başarı notu. Ancak örneklemde lise not ortalaması ve ACT puanı 0 olan öğrenci olmadığından yorumlanması anlamsız.
- ACT'ı sabit tutarak lise GPA notunu 1 puan artırdığımızda üniversite GPA'sı yarım puana yakın (0.453) artıyor. ACT notu aynı olan iki öğrenciden lise GPA'sı yüksek olanın üniversite GPA'sı da yüksek olacaktır.
- ACT'ın işareti $+$'dır ancak katsayısı çok küçük olduğu için etkisi fazla değil.

Örnek: Üniversite Başarısını Belirleyen Faktörler, gpa1.gdt

- Sadece ACT notunu alarak basit regresyon tahmin etseydik şöyle olacaktı:

$colGPA$ Basit Regresyon Tahmin Sonuçları

$$\widehat{colGPA} = 2.4 + 0.0271\ ACT$$

- ACT'ın katsayısı önceki çoklu regresyonda bulunandan 3 kat daha yüksek çıktı.
- Ancak, bu regresyon, bize, lise GPA'sı aynı iki öğrenciyi karşılaştırma olanağı vermiyor. Önceki regresyon veriyordu.
- Lise not ortalamasını kontrol ettiğimizde ACT puanının önemi azalıyor.

k Değişkenli Modelin Yorumu

SRF - Örneklem Regresyon Fonksiyonu

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k.$$

Değişimler Cinsinden

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 + \ldots + \hat{\beta}_k \Delta x_k.$$

- x_1 değişkeninin katsayısı $\hat{\beta}_1$'nın yorumu: diğer değişkenler sabitken, yani $\Delta x_2 = 0$, $\Delta x_3 = 0$, ..., $\Delta x_k = 0$

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1$$

- Diğer tüm değişkenler sabitken x_1'de meydana gelen 1 birim değişmenin y'de meydana getireceği ortalama değişim $\hat{\beta}_1$ kadardır (y'nin birimi cinsinden).
- Diğer katsayı tahminleri de benzer şekilde yorumlanır.

Örnek: Logaritmik Ücret Denklemi

Tahmin Sonuçları

$$\widehat{log\ Ucret} = 0.284 + 0.092\ egitim + 0.0041\ tecrube + 0.022\ kidem$$

$n = 526$ çalışan

- Katsayı tahminleri *ceteris paribus* yorumlanmalı.
- Bağımlı değişken logaritmik, açıklayıcı değişkenler kendi ölçü birimleriyle modelde yer aldığından (log-level) katsayı tahminleri 100 ile çarpılarak % olarak yorumlanmalı.
- Örneğin, tecrübe ve kıdem sabit tutulduğunda eğitim bir yıl arttırıldığında ücretler ortalama % 9.2 artmaktadır.
- Başka bir ifadeyle, tecrübe ve kıdem düzeyleri aynı olan iki çalışandan birinin eğitim düzeyi diğerinden bir yıl fazlaysa, bu iki çalışan için tahmin edilen ücret farkı ortalama % 9.2'dir.
- Burada somut iki işçiden değil ortalama durumdan bahsedilmektedir.

Diğer Değişkenleri Sabit Tutmanın Anlamı

- Çoklu regresyonda beta katsayılarını *ceteris paribus* koşulu altında bağımsız değişkenlerin y üzerindeki *kısmi etkileri* (partial effects) olarak yorumluyoruz.
- Örneğin, yukarıdaki regresyonda, $\hat{\beta}_1 = 0.092$, tecrübe ve kıdemi aynı olan iki işçiden eğitimi 1 yıl fazla olanının %9.2 daha yüksek ücret alacağı şeklinde yorumlandı.
- Bu yorum, verinin bu şekilde toplandığı anlamına gelmez. Veri (data) rassal seçilmiş 526 işçiye ait ücret, eğitim ve kıdem bilgilerinden oluşuyor. Kıdemi ve tecrübesi aynı olan işçileri ayrıca gruplandırmıyoruz.
- Aslında kıdemleri aynı olan işçilerden oluşan bir örneklem olsaydı kıdem değişkenini modele koymaya gerek kalmazdı.
- Ancak uygulamada çoğunlukla bu mümkün değildir. Çoklu regresyon analizinde zaten buna gerek yoktur.

Birden Fazla Değişkeni Aynı Anda Değiştirmek

- Bazen x'lerden birkaçını birden değiştirererk y'de meydana gelen değişimi ölçmek isteriz.
- Bazı durumlarda da x'lerden biri değiştirildiğinde diğeri otomatik olarak değişir.
- Örneğin ücret denkleminde kıdemi 1 yıl arttırdığımızda tecrübe de otomatik olarak 1 yıl artar.
- Bu durumda ikisinin ücret üzerindeki etkisi % 2.61 olur:

$$\Delta\widehat{log\ Ucret} = 0.0041\Delta tecrube + 0.022\Delta kidem$$

$$= 0.0041 + 0.022 = 0.0261$$

Tahmin Edilen Değerler ve Kalıntılar

$i$nci gözlem için tahmin edilen y değerleri (fitted/predicted values)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik}.$$

Kalıntılar (Residuals)

$$\hat{u}_i = y_i - \hat{y}_i$$

- x değerlerini tahmin edilen regresyon denkleminde yerine koyarsak modelce tahmin edilen y değerlerine ulaşırız.
- Gözlenen y değerleriyle modelce tahmin edilen değerler arasındaki fark kalıntıları verir.
- $\hat{u} > 0$ ise $y_i > \hat{y}_i$, eksik tahmin (underprediction)
- $\hat{u} < 0$ ise $y_i < \hat{y}_i$, fazla tahmin (overprediction)

Kalıntı Terimlerinin Cebirsel Özellikleri

- OLS kalıntılarının toplamı ve dolayısıyla da örnek ortalaması sıfıra eşittir:
$$\sum_{i=1}^{n} \hat{u}_i = 0, \qquad \bar{\hat{u}} = 0$$
Bu birinci örneklem moment koşulunun sonucudur.
- Açıklayıcı değişken x_j ile kalıntı terimleri arasındaki örneklem kovaryansı sıfırdır:
$$\sum_{i=1}^{n} x_{ij}\hat{u}_i = 0, \quad j = 1, 2, \ldots, k$$
Bu da diğer moment koşullarının bir sonucudur (bkz. OLS birinci sıra koşulları). Kalıntılarla açıklayıcı değişkenlerin ilişkisizliği empoze edilmiştir.
- $(\bar{x}_j, \bar{y} : j = 1, 2, \ldots, k)$ noktası daima OLS regresyon doğrusu üzerine düşer.
- $\bar{y} = \bar{\hat{y}}$

Katsayı Tahminlerinin Alternatif Türetimi

İki değişkenli model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2.$$

- Burada x_1'in eğim katsayısının tahmincisi:
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \hat{r}_{i1} y_i}{\sum_{i=1}^{n} \hat{r}_{i1}^2}$$
- Burada $\hat{r}_{i1}$, x_1'in x_2 üzerine regresyonundan elde edilen kalıntılardır:
$$x_{i1} = \hat{\alpha}_0 + \hat{\alpha}_1 x_{i2} + \hat{r}_{i1}$$
- Öyleyse $\hat{\beta}_1$, y'nin kalıntılar üzerine regresyonundan elde edilen eğim katsayısıdır:
$$y_i = \hat{\beta}_0 + \hat{\beta}_1 r_{i1} + kalıntı$$
- Ceteris paribus yorumunun başka bir versiyonu. (partialling out, netting out)

Basit ve Çoklu Regresyon Tahminlerinin Karşılaştırılması

Basit ve İki değişkenli model

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1, \quad vs. \quad \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2.$$

- Yukarıdaki regresyonlar genellikle farklı sonuçlar verir.
- Ancak şu iki durumda eğim katsayıları aynı olur:
- x_2'inin y üzerindeki kısmi etkisi sıfırdır, $\hat{\beta}_2 = 0$
- Örneklemde x_1 ve x_2 ilişkisizdir.

Kareler Toplamları (Sum of Squares)

- SST y'deki toplam değişkenliği verir.

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

  $\text{Var}(y) = SST/(n-1)$ olduğuna dikkat ediniz.
- Benzer şekilde SSE modelce açıklanan kısımdaki değişkenliği verir.

$$SSE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

- SSR ise kalıntılardaki değişkenliğin bir ölçütüdür.

$$SSR = \sum_{i=1}^{n}\hat{u}_i^2$$

- y'deki toplam değişkenlik aşağıdaki gibi yazılabilir:

$$SST = SSE + SSR$$

Uyum İyiliği (Goodness-of-fit)

- Bu ifadenin her iki tarafını SST'ye bölersek:

$$1 = \frac{SSE}{SST} + \frac{SSR}{SST}$$

- Açıklanan kısmın değişkenliğinin toplam değişkenlik içindeki payı regresyonun determinasyon (belirlilik) katsayısıdır ve R^2 ile gösterilir:

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

- SSE hiç bir zaman SST'den büyük olamayacağı için $0 \leq R^2 \leq 1$
- R^2 y'deki değişkenliğin x tarafından açıklanan kısmının yüzdesini verir. Regresyonun açıklama gücü yükseldikçe R^2 1'e yaklaşır.
- R^2 şu şekilde de hesaplanabilir: $R^2 = Corr(y, \hat{y})^2$

Uyum İyiliği (Goodness-of-fit)

- Determinasyon katsayısı:

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

- Regresyona yeni bir x eklendiğinde R^2 her zaman artar ya da aynı kalır.
- Bunun sebebi yeni bir değişken eklendiğinde SSR'nin her zaman azalmasıdır.
- Bu nedenle yeni bir değişkenin katkısının belirlenmesinde R^2 iyi bir ölçüt değildir.
- Bunun için düzeltilmiş R^2 (adjusted R^2) kullanılır.

Uyum İyiliği (Goodness-of-fit): Örnek

Basit ve İki değişkenli model

$$\widehat{colGPA} = 1.29 + 0.453\ hsGPA + 0.0094\ ACT$$

$$n = 141 \quad R^2 = 0.176$$

- Burada determinasyon katsayısı 0.176 olarak tahmin edilmiştir.
- Üniversite GPA notlarındaki değişkenliğin yaklaşık %17.6'sı $hsGPA$ ve ACT değişkenleriyle açıklanabilmektedir.
- Dışarıda bırakılan birçok faktör olduğundan üniversite başarısının küçük bir kısmı açıklanabilmiştir.
- Üniversite başarısını etkileyen bu modelde yer almayan başka birçok değişken olduğu unutulmamalıdır.

Orijinden Geçen Regresyon

x'ler 0 olduğunda tahmin edilen y değeri 0

$$\tilde{y} = \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \ldots + \tilde{\beta}_k x_k.$$

- Regresyonda sabit terim olmadığına dikkat ediniz.
- R^2 negatif çıkabilir. Bu durumda R^2 0 kabul edilir ya da regresyona sabit terim konarak yeniden tahmin yapılır.
- R^2'nin negatif çıkması, y'nin örneklem ortalamasının ($\bar{y}$) y'deki değişkenliği açıklamada modeldeki değişkenlerden daha başarılı olduğu anlamına gelir.
- Eğer PRF'de sabit terim sıfırdan farklı ise, orijinden geçen regresyonun OLS tahmincileri sapmalı olur.
- Sabit terim sıfır olduğu halde sıfır değilmiş gibi regresyona dahil etmek ise regresyonun varyansını yükseltir ve OLS tahmincilerinin değişkenliğini arttırır.

OLS Tahmincilerinin Sapmasızlığı için Gerekli Varsayımlar

MLR.1 Parametrelerde Doğrusallık

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

Model parametrelerde doğrusaldır. u rassal hata terimidir.

MLR.2 Rassal Örnekleme

Elimizde MLR.1 ile tanımlanan popülasyondan çekilmiş n gözlemli bir rassal örneklem vardır:

$$\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : \ i = 1, 2, \ldots, n\}$$

OLS Tahmincilerinin Sapmasızlığı için Gerekli Varsayımlar

MLR.3 Sıfır Koşullu Ortalama

$$\mathrm{E}(u|x_1, x_2, \ldots, x_k) = 0$$

- Bu varsayım x açıklayıcı değişkenlerinin kesin dışsal olduğunu söyler. Hata terimiyle açıklayıcı değişkenler ilişkisizdir.
- Bu varsayımın sağlanamadığı durumlar nelerdir?
- Bunlardan biri regresyonun fonksiyon kalıbının yanlış kurulmasıdır (functional form misspecification)
- Önemli bir değişkenin regresyon dışında bırakılması da bu varsayımı zedeler (omitted variable)
- Açıklayıcı değişkenlerde yapılan ölçme hataları bu varsayımın ihlaline yol açar (measurement error)
- Bu varsayım sağlanmıyorsa içsel değişkenler (endogenous variables) sözkonusudur.

OLS Tahmincilerinin Sapmasızlığı için Gerekli Varsayımlar

MLR.4 Tam Çoklu Bağıntının Olmaması

Bu varsayım x açıklayıcı değişkenleri arasında tam doğrusal bir ilişkinin olmaması gerektiğini söyler. Herhangi bir x diğer x'lerin lineer bir kombinasyonu olarak yazılamaz.

- Bu varsayım x'lerin birbirleriyle ilişkili olmasına izin verir. İzin verilmeyen tam korelasyonun olmamasıdır.
- x'ler tam ilişkili olursa OLS katsayılarının tahmini matematiksel olarak mümkün olmaz. (katsayılar belirsiz olur).
- Bu varsayıma göre açıklayıcı değişkenler ilişkili (correlated) olabilirler. x'ler arasında korelasyona izin vermezsek çoklu regresyondan istediğimiz faydayı alamayız.
- Örneğin, öğrenci notları, harcamaları ve aile geliri regresyonunda aile geliri (avginc) ile harcama (expend) arasında ilişki olduğunu bilerek bu değişkenleri modele sokuyoruz. Amaç geliri kontrol etmek.

OLS tahmincilerinin sapmasızlığı

TEOREM: $\hat{\beta}$'ların Sapmasızlığı

MLR.1-MLR.4 varsayımları altında OLS tahmin edicileri sapmasızdır:

$$\mathsf{E}(\hat{\beta}_j) = \beta_j, \quad j = 0, 1, 2, \ldots, k$$

Sapmasızlık OLS tahmin edicilerinin örnekleme dağılımlarının orta noktasının (beklentisinin) bilinmeyen popülasyon parametrelerine eşit olduğunu söyler.

Modele Gereksiz Açıklayıcı Değişken Eklenmesi

- Modele gerekli olmadığı halde bir açıklayıcı değişken eklersek OLS tahmini bundan nasıl etkilenir?
- Modele gereksiz bir değişken eklenmesi PRF'de bu değişkenin kısmi etkisinin sıfır olduğu anlamına gelmektedir. Model fazla kurulmuştur (overspecification).
- Örneğin aşağıdaki regresyonda x_3'ün kısmi etkisinin sıfır olduğunu varsayalım, $\beta_3 = 0$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

- Koşullu beklentisini alırsak

$$\mathsf{E}(y|x_1, x_2, x_3) = \mathsf{E}(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- Bu PRF'nin bilinmediğini, araştırmacının modele x_3'ü katsayısı 0 olduğu halde eklediğini varsayıyoruz.

Modele Gereksiz Açıklayıcı Değişken Eklenmesi

- Bu durumda SRF (Örneklem Regresyon Fonksiyonu - ÖRF)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

- OLS tahmin edicileri hala sapmasızdır:

$$\mathsf{E}(\hat{\beta}_0) = \beta_0, \quad \mathsf{E}(\hat{\beta}_1) = \beta_1, \quad \mathsf{E}(\hat{\beta}_2) = \beta_2, \quad \mathsf{E}(\hat{\beta}_3) = 0,$$

- Gereksiz eklenen değişkenin katsayısının doğru değeri 0'dır. Bu değişkenin bir açıklayıcılığı olmadığından OLS tahmincilerinin beklenen değeri de 0 olacaktır.
- OLS tahmincileri hala sapmasız olsa da regresyonun varyansı yükselir. Sonuç olarak tahmin edicilerin de varyansları (ve standart hataları) yüksek çıkacaktır.

Gerekli Bir Değişkenin Model Dışında Bırakılması (Omitted Variable)

- Modelde yer alması gerektiği halde bir değişkeni dışlarsak OLS tahmini bundan nasıl etkilenir?
- Gerekli bir değişkenin modelden dışlanması PRF'de bu değişkenin kısmi etkisinin sıfır olmadığı anlamına gelmektedir. Model eksik kurulmuştur (underspecification).
- Bu durumda OLS tahmin edicileri sapmalı olur.
- Örneğin MLR.1-MLR.4 varsayımları sağlansın ve PRF iki açıklayıcı değişken içersin:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

- x_2 değişkenini gözleyemediğimiz için model dışında bıraktığımızı düşünelim. Bu durumda SRF

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$

- x_1 katsayı tahmincisi $\tilde{\beta}_1$ hala sapmasız mıdır?

Gerekli Bir Değişkenin Model Dışında Bırakılması (Omitted Variable)

- Dışarıda bırakılan değişkenin etkisi hata teriminin içinde yer alacaktır:
$$y = \beta_0 + \beta_1 x_1 + \nu$$
- Gerçek model (PRF) x_2'yi içermektedir. Bu nedenle hata terimi ν
$$\nu = \beta_2 x_2 + u$$
- Yukarıdaki modelde β_1'in OLS tahmincisi:
$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)y_i}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}$$
- Sapmanın boyutunu belirlemek için $\tilde{\beta}_1$ formülünde y yerine PRF'yi yazıp, yeniden düzenleyerek beklentisini alıyoruz.

Gerekli Bir Değişkenin Model Dışında Bırakılması

$$\begin{aligned}\tilde{\beta}_1 &= \frac{\sum(x_{i1} - \bar{x}_1)(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i)}{\sum(x_{i1} - \bar{x}_1)^2} \\ &= \beta_1 + \beta_2 \frac{\sum(x_{i1} - \bar{x}_1)x_{i2}}{\sum(x_{i1} - \bar{x}_1)^2} + \frac{\sum(x_{i1} - \bar{x}_1)u_i}{\sum(x_{i1} - \bar{x}_1)^2}\end{aligned}$$

Bunun beklenen değerini (koşullu) alırsak

$$\begin{aligned}E(\tilde{\beta}_1) &= \beta_1 + \beta_2 \frac{\sum(x_{i1} - \bar{x}_1)x_{i2}}{\sum(x_{i1} - \bar{x}_1)^2} + \frac{\sum(x_{i1} - \bar{x}_1)\overbrace{\mathsf{E}(u_i)}^{=0, MLR.3}}{\sum(x_{i1} - \bar{x}_1)^2} \\ &= \beta_1 + \beta_2 \left(\frac{\sum(x_{i1} - \bar{x}_1)x_{i2}}{\sum(x_{i1} - \bar{x}_1)^2} \right)\end{aligned}$$

Gerekli Bir Değişkenin Model Dışında Bırakılması

$$E(\tilde{\beta}_1) = \beta_1 + \beta_2 \left(\frac{\sum(x_{i1} - \bar{x}_1)x_{i2}}{\sum(x_{i1} - \bar{x}_1)^2} \right)$$

β_2'nin sağında yer alan parantez içindeki ifade x_2'nin x_1 üzerine regresyonundan elde edilen eğim katsayısıdır:

$$\tilde{x}_2 = \tilde{\delta}_0 + \tilde{\delta}_1 x_1$$

Böylece

$$E(\tilde{\beta}_1) = \beta_1 + \beta_2 \tilde{\delta}_1$$

$$sapma = E(\tilde{\beta}_1) - \beta_1 = \beta_2 \tilde{\delta}_1$$

olur. Buna **dışlanmış değişken sapması** (omitted variable bias) adı verilir.

Dışlanmış Değişken Sapması

$$sapma = E(\tilde{\beta}_1) - \beta_1 = \beta_2 \tilde{\delta}_1$$

- Eğer $\tilde{\delta}_1 = 0$ ya da $\beta_2 = 0$ ise sapma 0 olur.
- Sapmanın işareti hem β_2'ye hem de dışlanan değişken ile modele dahil edilen değişken arasındaki korelasyona bağlıdır.
- Dışlanan değişken gözlenemiyor ise bu korelasyonu hesaplamak mümkün olmayabilir.
- Aşağıdaki tablo sapmanın yönüne ilişkin olası durumları özetlemektedir:

Sapmanın İşareti

	$Corr(x_1, x_2) > 0$	$Corr(x_1, x_2) < 0$
$\beta_2 > 0$	pozitif sapma	negatif sapma
$\beta_2 < 0$	negatif sapma	pozitif sapma

Dışlanmış Değişken Sapması

$$sapma = E(\tilde{\beta}_1) - \beta_1 = \beta_2 \tilde{\delta}_1$$

- Sapmanın işaretinin yanı sıra boyutu da önemlidir. Sapmanın boyutu hem $\tilde{\delta}_1$'e hem de β_2'ye bağlıdır.
- β_1'in büyüklüğüne nazaran küçük bir sapma uygulamada sorun yaratmayabilir.
- Ancak çoğu durumda sapmanın büyüklüğünü hesaplamak mümkün olmaz.
- Bazı durumlarda sapmanın yönü hakkında bir fikir edinebiliriz. Örneğin ücret denkleminde gerçek PRF hem eğitim (educ) hem de doğuştan gelen yetenek (ability) değişkenlerini içersin.
- Yetenek (ability) değişkeni gözlenemediği için model dışında bırakılırsa dışlanmış değişken sapması oluşur.
- Bu durumda eğitim katsayısındaki sapmanın işaretinin $+$ olacağının söyleyebiliriz. Çünkü, yetenekli insanlar daha fazla eğitim alma eğilimindedir ve yetenek ücretlerle pozitif ilişkilidir.

Dışlanmış Değişken Sapması

- Dışlanmış değişkenin etkisi hata terimi u'nun içinde yer alacağı için MLR.3 artık sağlanmaz.

$$wage = \beta_0 + \beta_1 educ + \beta_2 ability + u$$

- Bunun yerine aşağıdaki model tahmin edilirse

$$wage = \beta_0 + \beta_1 educ + \nu$$

$$\nu = \beta_2 ability + u$$

- Eğitim değişkeni ile hata terimi (ν) ilişkili olur.
- Bu durumda MLR.3 sağlanmaz:

$$\mathrm{E}(\nu | educ) \neq 0$$

- Eğitim değişkeni içseldir. Yeteneğin dışlandığı durumda eğitimin etkisi abartılı tahmin edilir. Aslında eğitimin etkisinin bir kısmı doğuştan gelen yeteneğe bağlıdır.

Dışlanmış Değişken Sapması

- Daha fazla açıklayıcı değişkenin içerildiği modellerde gerekli bir değişkenin model dışında bırakılması OLS tahmincilerinin genellikle sapmalı olmasına yol açar.
- Gerçek model aşağıdaki gibi olsun:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

- x_3 dışarıda bırakılarak aşağıdaki model tahmin edilsin:

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2$$

- x_3, x_1 ile ilişkili, x_2 ile ilişkisiz olsun.
- Bu durumda $\tilde{\beta}_1$ ve $\tilde{\beta}_2$'nin her ikisi birden sapmalı olacaktır. Ancak x_1 ile x_2 ilişkisiz ise $\tilde{\beta}_2$ sapmasız olacaktır.

OLS Tahmincilerinin Varyansları

MLR.5 Sabit Varyans (Homoscedasticity)

Bu varsayım x açıklayıcı değişkenlerine koşullu olarak hata varyansının sabit olduğunu söyler:

$$\mathrm{Var}(u | x_1, x_2, \ldots, x_k) = \sigma^2$$

- Bunun sağlanmadığı duruma **değişen varyans** (heteroscedasticity) denir.
- Bu varsayım OLS tahmincilerinin varyanslarının ve standart hatalarının türetilmesinde ve etkinlik özelliklerinin belirlenmesinde kullanılır.
- Sapmasızlık için bu varsayıma gerek yoktur.
- Örneğin, ücret denkleminde bu varsayım, model dışında bırakılan faktörlerin değişkenliğinin modele dahil edilen değişkenlere (tecrübe, eğitim, kıdem, vs.) bağlı olmadığını söylemektedir.

OLS Tahmincilerinin Varyansları

Gauss-Markov Varsayımları

MLR.1-MLR.5 varsayımlarına *Gauss-Markov Varsayımları* denir.
MLR.1: Parametrelerde doğrusallık,
MLR.2: Rassal örnekleme,
MLR.3: Sıfır koşullu ortalama,
MLR.4: Tam çoklu doğrusallığın olmaması,
MLR.5: Sabit varyans.

- Bu varsayımlar kesit-veri regresyonu için geçerli varsayımlardır.
- Zaman serileriyle regresyon analizinde bu varsayımların değiştirilmesi gerekir.
- MLR.3 ve MLR.5 bağımlı değişken cinsinden ifade edilebilir:

$$\mathrm{E}(y|x_1, x_2, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$
$$\mathrm{Var}(y|x_1, x_2, \ldots, x_k) = \mathrm{Var}(u|x_1, x_2, \ldots, x_k) = \sigma^2$$

OLS Tahmincilerinin Varyansları

Teorem: $\hat{\beta}$'ların varyansları

Gauss-Markov varsayımları (MLR.1-MLR.5) altında

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}, \quad j = 1, 2, \ldots, k$$

Burada

$$SST_j = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

x_j'deki örneklem değişkenliği, R_j^2 ise x_j'nin diğer tüm x değişkenlerine (sabit terim içeren) regresyonundan elde edilen belirlilik katsayısıdır.

- $\mathrm{Var}(\hat{\beta}_j)$, σ_2 ile aynı yönde, SST_j ile ters yönde ilişkilidir.
- SST_j'yi arttırmanın tek yolu gözlem hacmini (n) arttırmaktır. σ^2'yi düşürmenin tek yolu ise güçlü açıklayıcı değişkenler bulmaktır.

OLS Tahmincilerinin Varyansları

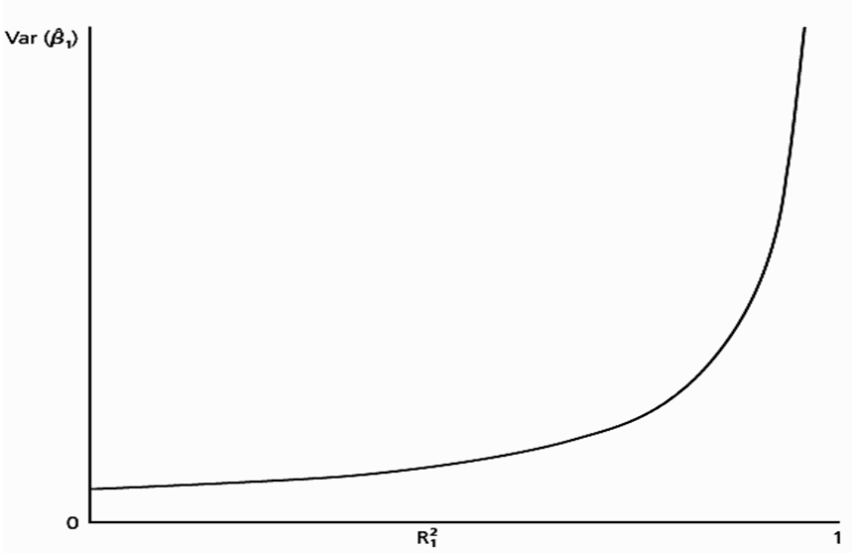
Teorem: $\hat{\beta}$'ların varyansları

Gauss-Markov varsayımları (MLR.1-MLR.5) altında

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}, \quad j = 1, 2, \ldots, k$$

- $\mathrm{Var}(\hat{\beta}_j)$, x'lerin birbirleriyle korelasyon düzeyini belirten R_j^2 terimine de bağlıdır.
- Tek açıklayıcı değişkenli modelde bu terim bulunmaz.
- Açıklayıcı değişkenlerin birbileriyle doğrusal ilişki düzeyi arttıkça OLS tahmincilerinin varyansı sınırsız artar.
- x'ler arasında yüksek doğrusal bağlantı (multicollinearity) olması varyansların yüksek çıkmasına neden olur.
- Limitte $R_j^2 = 1$ olduğunda varyans sonsuz olur (ayrıca $\hat{\beta}$'lar belirsiz olur). Ancak MLR.4 varsayımı bunu engeller.

Varyans ve R_j^2 İlişkisi



OLS Tahmincilerinin Varyansları

Varyansın Tahmini

Hata varyansının sapmasız bir tahmincisi şudur:

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{SSR}{n-k-1}$$

- Serbestlik derecesi (degrees of freedom):

$$dof = n - (k+1)$$

- dof = gözlem sayısı − parametre sayısı
- Serbestlik derecesi OLS birinci sıra koşullarından gelmektedir. Bu koşullar $k+1$ taneydi. Kalıntılar üzerine $k+1$ tane kısıt konmaktadır.
- n tane kalıntı teriminden $n-(k+1)$ tanesi biliniyorsa geriye kalan $k+1$ kalıntı otomatik olarak bilinecektir. Öyleyse kalıntıların serbestlik derecesi $n-k-1$'dir.
- Hata terimi u'nun serbestlik derecesi ise n'dir.

OLS Tahmincilerinin Varyansları

$\hat{\beta}$'ların Standart Sapmaları (sd)

$$\text{sd}(\hat{\beta}_j) = \frac{\sigma}{\sqrt{SST_j(1-R_j^2)}}, \quad j = 1, 2, \ldots, k$$

$\hat{\beta}$'ların Standart Hataları (se)

$$\text{se}(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{SST_j(1-R_j^2)}}, \quad j = 1, 2, \ldots, k$$

- $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ regresyonun standart hatasıdır (SER - standard error of regression).
- SER regresyonun standart sapmasının bir tahmincisidir. Regresyona yeni bir değişken eklendiğinde SER azalabilir ya da artabilir.
- $\text{se}(\hat{\beta}_j)$ güven aralıklarının hesaplanmasında ve hipotez testlerinin yapılmasında kullanılır.

Gauss-Markov Teoremi

Gauss-Markov Teoremi

MLR.1-MLR.5 varsayımları altında sıradan en küçük kareler (OLS) tahmin edicileri, tüm doğrusal, sapmasız tahmin ediciler kümesi içinde en etkin (en küçük varyanslı) olanlarıdır. Başka bir ifadeyle, MLR.1-MLR.5 varsayımları altında OLS tahmin edicileri $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$; populasyon parametreleri $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$'nin Doğrusal En İyi Sapmasız Tahmin Edicileridir (kısaca, DESTE ya da BLUE-Best Linear Unbiased Estimators)

- Gauss-Markov teoremi regresyon modelinin OLS yöntemiyle tahmini için teorik dayanak sağlar.
- Eğer bu varsayımlar sağlanıyorsa OLS dışında başka bir tahmin yöntemine başvurmamıza gerek yoktur. OLS bize varyansı en düşük (best) tahmincileri vermektedir.
- Bu 5 varsayımdan biri bile ihlal edilirse Gauss-Markov teoremi geçersiz olur. MLR.3 sağlanmazsa sapmasızlık, MLR.5 sağlanmazsa etkinlik özelliği kaybolur.

OLS Tahmincilerinin Doğrusallığı

Doğrusal tahmin ediciler

$\tilde{\beta}_j$ tahmincisi aşağıdaki gibi yazılabiliyorsa doğrusaldır:

$$\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i$$

Burada w_{ij} tüm açıklayıcı değişkenlerin bir fonksiyonu olabilir.

OLS tahmincileri yukarıdaki gibi yazılabildiğinden doğrusaldırlar:

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n}\hat{r}_{ij}y_i}{\sum_{i=1}^{n}\hat{r}_{ij}^2} = \sum_{i=1}^{n} w_{ij} y_i, \quad \text{burada} \quad w_{ij} = \frac{\hat{r}_{ij}}{\sum_{i=1}^{n}\hat{r}_{ij}^2}$$

$\hat{r}_{ij}$ x_j'nin tüm diğer açıklayıcı değişkenler üzerine regresyonundan elde edilen kalıntı terimidir.