ÇOKLU REGRESYON ANALİZİNDE EK KONULAR

Hüseyin Taştan[1]

[1]Yıldız Teknik Üniversitesi
İktisat Bölümü

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

14 Ekim 2012

Regresyon Analizi: Ek Konular

Bu bölümde aşağıdaki konuları inceleyeceğiz

- Veri ölçeğinin (data scaling) tahminlere etkisi
- Standartlaştırılmış regresyon
- Fonksiyonel kalıp ile ilgili ek konular: Karesel (quadratic) modeller, Etkileşim terimli (interaction term) modeller
- Regresyonda uyumun iyiliği ölçütleri ve değişkenlerin seçimi: Düzeltilmiş R^2
- Kestirim (Prediction)

Ölçü Birimlerinin Tahmin Sonuçlarına Etkisi

- Değişkenlerin ölçü birimlerini değiştirmek, katsayılardaki fazla sıfırları yok etmek gibi "*regresyonun görünümünü iyileştirmek ve yorumunu kolaylaştırmak*" amacıyla yapılır.
- Regresyonun özünü değiştirmez, tüm test sonuçları ve bulgular aynı kalır.
- Değişkenlerin ölçü birimini değiştirmek betaların anlamlılık düzeyini etkilemez. t istatistikleri aynı kalır.
- Değişkenlerin ölçü biriminin değişmesi determinasyon katsayısı R^2'yi etkilemez.
- Buna karşılık SSR ve SER ölçü birimlerine göre değişir.
- Değişkenlerin ölçü biriminin değişmesi F testini etkilemez.

Standartlaştırılmış Regresyon

- x_j, 1 birim değil de 1 standart sapma değişseydi y ne kadar değişirdi?
- Bu soruyu yanıtlayabilmek için regresyondaki tüm değişkenleri (y ve tüm x'ler) standart hale getirip sonra bu standartlaştırılmış değişkenlerle regresyon tahmin etmemiz gerekir.
- Bir değişkeni, kendi (aritmetik) ortalamasından farkını alıp (örneklem) standart sapmasına bölersek, o değişkeni standartlaştırmış oluruz:

$$z_y = \frac{y - \bar{y}}{\hat{\sigma}_y},$$

$$z_1 = \frac{x_1 - \bar{x}_1}{\hat{\sigma}_1}, z_2 = \frac{x_2 - \bar{x}_2}{\hat{\sigma}_2}, \ldots, z_k = \frac{x_k - \bar{x}_k}{\hat{\sigma}_k},$$

- $\hat{\sigma}_j$, x_j'nin örneklem standart sapmasıdır.

Standartlaştırılmış Regresyon

Aşağıdaki OLS örneklem regresyon fonksiyonunu standartlaştırmak istiyoruz:

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

Her bir değişkenin ortalamasından farkını alarak modeli yeniden yazarsak:

$$y_i - \bar{y} = \hat{\beta}_1(x_{i1} - \bar{x}_1) + \hat{\beta}_2(x_{i2} - \bar{x}_2) + \ldots + \hat{\beta}_k(x_{ik} - \bar{x}_k) + \hat{u}_i$$

$\hat{u}$'nın ortalaması sıfırdır. Modelde sabit terim yoktur. Eşitliğin her iki tarafında yer alan terimleri değişkenlerin örneklem standart sapmalarına bölersek aşağıdaki modele ulaşırız:

$$\frac{y_i - \bar{y}}{\hat{\sigma}_y} = \frac{\hat{\sigma}_1}{\hat{\sigma}_y}\hat{\beta}_1\frac{(x_{i1} - \bar{x}_1)}{\hat{\sigma}_1} + \frac{\hat{\sigma}_2}{\hat{\sigma}_y}\hat{\beta}_2\frac{(x_{i2} - \bar{x}_2)}{\hat{\sigma}_2} + \ldots + \frac{\hat{\sigma}_k}{\hat{\sigma}_y}\hat{\beta}_k\frac{(x_{ik} - \bar{x}_k)}{\hat{\sigma}_k} + \frac{\hat{u}_i}{\hat{\sigma}_y}$$

Standartlaştırılmış Regresyon

- Modeli yeniden yazarsak:

$$z_y = \hat{b}_1 z_1 + \hat{b}_2 z_2 + \ldots + \hat{b}_k z_k + hata$$

Burada

$$z_y = \frac{y - \bar{y}}{\hat{\sigma}_y}, \quad z_j = \frac{x_j - \bar{x}_j}{\hat{\sigma}_j}, \ j = 1, 2, \ldots, k$$

- Eğim katsayıları: standardize edilmiş katsayıları, ya da beta katsayıları

$$\hat{b}_j = \frac{\hat{\sigma}_j}{\hat{\sigma}_y}\hat{\beta}_j, \ j = 1, 2, \ldots, k$$

- Yorum: x_j'de meydana gelen 1 standart sapmalık değişime karşılık y'de meydana gelen değişim $\hat{b}_j$ standart sapma kadardır.
- Kısmi etkiler artık ölçü birimlerinden bağımsızdır ve birbirleriyle karşılaştırılabilir.

Standartlaştırılmış Regresyon: Örnek

Hava Kirliliği ve Ev Fiyatları: hprice2.gdt

Bağımlı değişken: o bölgedeki evlerin medyan fiyatının logaritması(log(price))
Açıklayıcı değişkenler:
nox: bölgedeki hava kirliliği ölçütü,
dist: bölgenin iş merkezlerine uzaklığı,
crime: bölgedeki kişi başına suç sayısı,
rooms: bölgedeki evlerin ortalama oda sayısı,
stratio: ortalama öğrenci-öğretmen oranı

Seviyelerle model:

$$price = \beta_0 + \beta_1 nox + \beta_2 crime + \beta_3 rooms + \beta_4 dist + \beta_5 stratio + u$$

Standartlaştırılmış model:

$$zprice = b_1 znox + b_2 zcrime + b_3 zrooms + b_4 zdist + b_5 zstratio + zu$$

Standartlaştırılmış Regresyon: Örnek

Standartlaştırılmış model tahmin sonuçları

$$\widehat{zprice} = -0.340\ znox - 0.143\ zcrime + 0.514\ zrooms - 0.235\ zdist - 0.270\ zstratio$$

- Hava kirliliğinde (nox) 1 standart sapma artış ev fiyatlarını 0.34 standart sapma azaltmaktadır.
- Suç oranındaki 1 standart sapma artış ev fiyatlarını 0.143 standart sapma azaltmaktadır.
- Hava kirliliği göreceli olarak ev fiyatları üzerinde suç oranından daha büyük etkiye sahiptir.
- Oda sayısı en yüksek standartlaştırılmış etkiye sahip değişkendir.
- Açıklayıcı değişkenlerin medyan ev fiyatları üzerindeki para birimi cinsinden etkilerini görmek istiyorsak standartlaştırılmamış regresyonu tahmin etmemiz gerekir.

Örnek, Standartlaştırılmamış Model Tahmin Sonuçları

$$\widehat{\text{price}} = \underset{(5054.6)}{20871.1} - \underset{(354.09)}{2706.43}\,\text{nox} - \underset{(32.929)}{153.601}\,\text{crime} + \underset{(393.60)}{6735.50}\,\text{rooms}$$
$$- \underset{(188.11)}{1026.81}\,\text{dist} - \underset{(127.43)}{1149.20}\,\text{stratio}$$
$$n = 506 \quad \bar{R}^2 = 0.6320 \quad F(5, 500) = 174.47 \quad \hat{\sigma} = 5586.2$$

(standard errors in parentheses)

Regresyonun Fonksiyonel Biçimi

- Daha önce, bağımlı ve/veya bağımsız değişkenleri doğal logaritma cinsinden ifade ederek regresyonda doğrusal-olmayan ilişkilerin yakalanabileceğini görmüştük.
- Örnek: Ev fiyatları modeli

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_3 rooms + u$$

- β_1: Ev fiyatlarının hava kirliliğine göre esnekliği
- $100\beta_2$: Oda sayısında 1 artışın ev fiyatlarında yol açacağı % değişim, yarı-esneklik (semi-elasticity)

Örnek

Ev fiyatları modeli

$$\widehat{\log(\text{price})} = \underset{(0.188)}{9.234} - \underset{(0.066)}{0.718}\,\log(\text{nox}) + \underset{(0.019)}{0.306}\,\text{rooms}$$
$$n = 506 \quad \bar{R}^2 = 0.512 \quad F(2, 503) = 265.69 \quad \hat{\sigma} = 0.28596$$

(standard errors in parentheses)

- Oda sayısı sabitken, hava kirliliği ölçütündeki (nox) %1 artış ev fiyatlarını ortalamada % 0.718 azaltmaktadır. Hava kirliliği esnekliği 0.718'dir.
- Hava kirliliği sabitken, oda sayısındaki 1 artış ev fiyatlarını ortalamada %30.6 (100×0.306) arttırmaktadır.
- $\log(y)$'deki değişme büyüdükçe $\%\Delta y \approx 100 \times \Delta \log(y)$ yaklaştırımı bozulur. Sonuç olarak büyük oranda yaklaştırım hatası ortaya çıkabilir.

Yaklaştırım Hatası

- Logaritmik bağımlı değişkendeki büyük değişimler için aşağıdaki formül kullanılabilir:

$$\widehat{\%\Delta y} = 100 \times [\exp(\hat{\beta}_2) - 1]$$

- Önceki örnekte $\hat{\beta}_2 = 0.306$ bulunmuştu:

$$\widehat{\%\Delta y} = 100 \times [\exp(0.306) - 1] = \%35.8$$

- Yarı-esneklik daha yüksek bulundu.
- $\exp(\hat{\beta}_2)$ sapmalı ancak tutarlı bir tahmincidir (neden?)

Doğal Log Dönüştürmesinin Avantajları

- Negatif olmayan değerler alan bir bağımlı değişkeni ($y > 0$) logaritmik olarak ifade etmek pek çok avantaj sağlar.
- Katsayılar x'lerin ölçü birimlerinden bağımsız olarak, esneklik ya da yarı-esneklik şeklinde tahmin edilir.
- $y > 0$ iken, $\log(y)$, CLM varsayımlarının sağlanması açısından, y serisine kıyasla çok daha elverişlidir. $y > 0$ düzey (level) değişkeni genellikle değişen varyanslı (heteroscedastic) ve çarpık (skewed) bir koşullu dağılıma sahiptir.
- Logaritma dönüştürmesi çarpıklığı azaltır ve varyansdaki değişmeyi yumuşatır.
- Log alınması değişkenin aralığını (range) büyük ölçüde düşürür. Bu ise, tahmin edicilerin aşırı uç değerlerden (outliers) fazla etkilenmemesini sağlar.

Doğal Log Dönüştürmesinin Avantajları

- Ücret, gelir, nüfus, üretim, satışlar vb gibi pozitif değerler alan değişkenleri regresyona genellikle düzey (level) olarak değil logaritmik olarak ekleriz.
- İşsizlik oranı, faiz oranı, herhangi bir projeye vs katılma oranı gibi oranları genellikle düzey olarak regresyona dahil ederiz. Ancak, her gözlemi pozitif olan oranların bazen Log biçiminde regresyona sokulduğu da görülmektedir.
- Oranlar (işsizlik oranı, örneğin) düzey olarak alınmışsa, yorum yaparken, "işsizlik oranında bir birimlik (= yüzde 1 puanlık – **a percentage point increase**) artış olduğunda" deriz.
- Oran Log olarak (Log(işsizlik oranı)) alınmışsa, "işsizlik oranında %1'lik (**a percentage increase**) artış olduğunda" diye yorumlarız.
- İşsizlik oranı düzey olarak %8 den %9'a yükselmişse, artış %1 puandır. Ama yüzde artış olarak Log(9)-Log(8)=0.1177=%11.77'lik bir artış olmuştur. İkisinin birbirine karıştırılmaması gerekir.

Doğal Log Dönüştürmesinin Avantajları

- Seri negatif olmayan değerler alıyorsa (≥ 0), yani pozitif sayıların yanında bazı gözlemler sıfır değerini de alıyorsa Log kullanamayız, zira $Log(0)$ tanımlanamaz.
- Bu halde, y serisini $\log(y)$'ye çeviremeyiz, ancak, $\log(1+y)$ serisini $\log(y)$ yerine kullanabiliriz.
- Eğer seride 0 değeri seyrek ise bu yola başvurabiliriz. Bu durumda katsayıların yorumu yine $\log(y)$ kullanıldığındaki gibidir. Büyük bir fark oluşmamaktadır.
- Bağımlı değişkenleri $\log(y)$ ve y olan iki regresyonun R^2'leri doğrudan karşılaştırılamaz. Gerekli dönüşürme işlemlerinin yapılması gerekir.

Fonksiyon Kalıbı: Karesel Modeller

- Değişkenlerin marjinal etkileri sabit değil de artan ya da azalan türde ise karesel modeller kullanmalıyız.
- Bu durumda eğim katsayısı sabit değildir. x'in hangi değeri aldığına bağlıdır.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$$

- x ile y ilişkisindeki eğim aşağıdaki gibi yaklaştırılabilir:

$$\Delta y \approx (\hat{\beta}_1 + 2\hat{\beta}_2 x)\Delta x$$

- Ya da

$$\frac{\Delta y}{\Delta x} \approx (\hat{\beta}_1 + 2\hat{\beta}_2 x)$$

- $x = 0$ ise $\hat{\beta}_1$, x 0'dan 1'e değişirken eğim katsayısı tahminini verir. $x = 1$ ve daha yüksek değerler için ikinci terimin dikkate alınması gerekir.

Karesel Modeller: Örnek

$$wage = 3.73 + 0.298\ exper - 0.0061\ exper^2$$

- $\beta_1 > 0, \beta_2 < 0$ ise $\frown$ şeklinde ilişki olur.
- $\beta_1 < 0, \beta_2 > 0$ ise ilişki $\smile$ şeklindedir.
- Yukarıdaki regresyon tecrübenin ücretler üzerinde azalan bir etkiye sahip olduğunu gösteriyor.
- Tahmini eğim:

$$\frac{\Delta \widehat{wage}}{\Delta exper} \approx 0.298 - (2 \times 0.0061) exper$$

- İlk bir yıllık tecrübe ücretlerde 0.298 dolarlık bir artış yaratırken, ikinci yıldaki eğim

$$\frac{\Delta \widehat{wage}}{\Delta exper} = 0.298 - 0.0122(1) = 0.286$$

Karesel Modeller: Örnek

$$wage = 3.73 + 0.298\ exper - 0.0061\ exper^2$$

- Tecrübe 10. yıldan 11. yıla değişirken:

$$\frac{\Delta \widehat{wage}}{\Delta exper} = 0.298 - 0.0122(10) = 0.176$$

- Dönüm noktası:

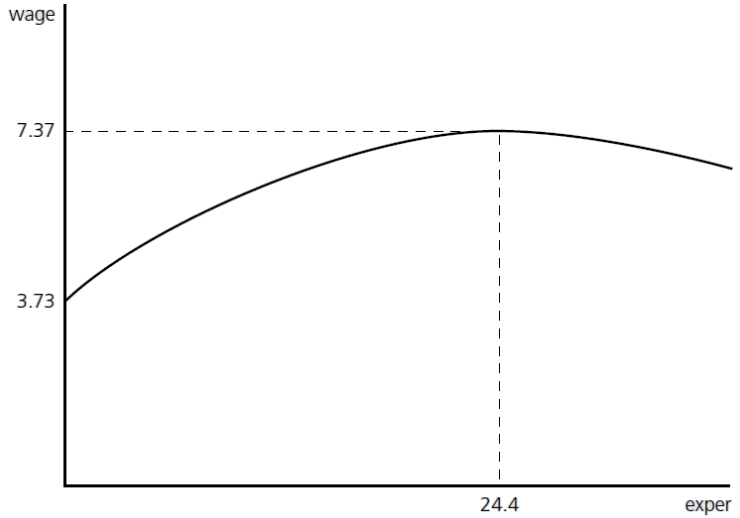
$$\frac{\Delta \hat{y}}{\Delta x} \approx (\hat{\beta}_1 + 2\hat{\beta}_2 x) = 0 \Rightarrow x^* = \frac{\hat{\beta}_1}{2\hat{\beta}_2}$$

- Ücret-tecrübe ilişkisinde dönüm noktası:

$$exper^* = 0.298/0.0122 = 24.4$$

Karesel Modeller: Ücret-tecrübe ilişkisi

$$wage = 3.73 + 0.298\ exper - 0.0061\ exper^2$$



Karesel Modeller: Örnek

$$\widehat{\log(\text{price})} = \underset{(0.566)}{13.386} - \underset{(0.115)}{0.902}\log(\text{nox}) - \underset{(0.043)}{0.0868}\log(\text{dist}) - \underset{(0.0059)}{0.0476}\,\text{stratio} - \underset{(0.1655)}{0.5451}\,\text{rooms} + \underset{(0.0128)}{0.0623}\,\text{rooms}^2$$

$$n = 506 \quad \bar{R}^2 = 0.5988 \quad F(5, 500) = 151.77 \quad \hat{\sigma} = 0.25921$$

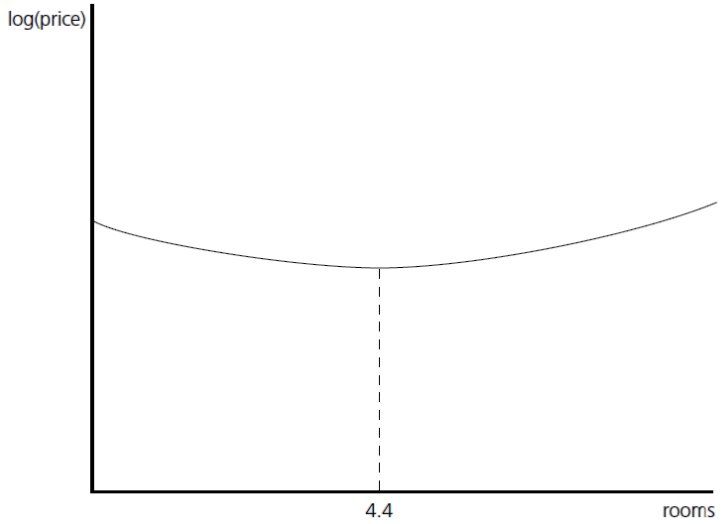
- Burada evin değeri ile oda sayısı arasında önce azalan sonra artan bir ilişki vardır.
- Ev oda sayısı 3'den 4'e değişirken fiyatlarda oluşan etki:

$$\frac{\Delta \widehat{\log(price)}}{\Delta rooms} = -0.5451 + 0.1246(3) = -0.1713 \approx -\%17.13$$

- Rooms=3 iken fazladan bir odanın fiyatlarda yol açacağı azalma yaklaşık %17.13'dür.
- Dönüm noktası:

$$rooms^* = 0.5451/0.1246 = 4.37 \approx 4.4$$

Karesel Modeller: Ev fiyatları



Karesel Modeller: Örnek

- İlave bir odanın fiyatlarda yaratacağı değişme:

$$\Delta\widehat{\log(price)} = [-0.545 + 2(0.062)rooms]\Delta rooms$$

$$\%\Delta\widehat{price} = 100 \times [-0.545 + 2(0.062)rooms]\Delta rooms$$
$$= (-54.5 + 12.4rooms)\Delta rooms$$

- Örneğin oda sayısı 5'den 6'ya değişirken oluşan etki yaklaşık $-54.5 + 12.4 \times 5 = \%7.5$. Burada $\Delta rooms = 1$ olduğuna dikkat ediniz.
- Oda sayısı 6'den 7'ye değişirken oluşan etki yaklaşık $-54.5 + 12.4 \times 6 = \%19.9$.
- Oda sayısı 5'den 7'ye değişirken oluşan etki yaklaşık $(-54.5 + 12.4 \times 5)2 = \%15$. Bu durumda $\Delta rooms = 2$.

Karşılıklı Etkileşim (Interaction) Terimi İçeren Modeller

- Bazı durumlarda bir açıklayıcı değişkenin kısmi etkisi başka bir açıklayıcı değişkenin aldığı değere bağlı olabilir.
- Bunu yakalamak için modele etkileşim terimleri eklenebilir:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \underbrace{x_1 \times x_2}_{interaction} + \beta_4 x_3 + u$$

- Yukarıdaki modelde x_1 ve x_2 değişkenleri karşılıklı etkileşime sahiptir. x_1'in y üzerindeki kısmi etkisi x_2'nin değerine bağlıdır:

$$\frac{\Delta y}{\Delta x_1} = \beta_1 + \beta_3 x_2$$

- Bu kısmi etkinin hesaplanabilmesi için x_2 yerine bir değer girilmesi gerekir. Genellikle x_2'nin ortalaması, medyanı gibi değerler kullanılır.
- Benzer şekilde x_2'nin kısmi etkisi x_1'e bağlıdır:

$$\frac{\Delta y}{\Delta x_2} = \beta_2 + \beta_3 x_1$$

Karşılıklı Etkileşim (Interaction) Terimi İçeren Modeller

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \underbrace{x_1 \times x_2}_{interaction} + \beta_4 x_3 + u$$

$$\frac{\Delta y}{\Delta x_1} = \beta_1 + \beta_3 x_2$$

- x_2'nin örneklem ortalaması $\bar{x}_2$ olsun. Kısmi etkide yerine yazarsak

$$\frac{\Delta y}{\Delta x_1} = \beta_1 + \beta_3 \bar{x}_2$$

- Bu bize $x_2 = \bar{x}_2$ iken kısmi etkiyi verir. Peki bu kısmi etki istatistik bakımından anlamlı mı?
- Bunu test etmek için $x_1 \times x_2$ etkileşimi yerine $x_1 \times (x_2 - \bar{x}_2)$ etkileşimini yazarak modeli yeniden tahmin edeceğiz:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \underbrace{x_1 \times (x_2 - \bar{x}_2)}_{interaction} + \beta_4 x_3 + u$$

Karşılıklı Etkileşim (Interaction) Terimi İçeren Modeller

- Bunu test etmek için $x_1 \times x_2$ etkileşimi yerine $x_1 \times (x_2 - \bar{x}_2)$ etkileşimini yazarak modeli yeniden tahmin edeceğiz:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \underbrace{x_1 \times (x_2 - \bar{x}_2)}_{interaction} + \beta_4 x_3 + u$$

- Bu regresyonda x_1'in katsayısının t testi kısmi etkinin anlamlı olup olmadığını gösterir.

$$H_0 : \beta_1 = 0$$

- Diğer değişkenlerin kısmi etkilerinin anlamlılığı da benzer şekilde sınanabilir.

Etkileşim Terimli Modeller: Örnek, attend.gdt

Değişken tanımları:
stndfnl: standardize edilmiş öğrenci final notu; **atndrte**:devamlılık oranı (%); **priGPA**: önceki sınıfların ortalama notu (4 üzerinden); **ACT**: genel yetenek sınav notu,

$$\widehat{stndfnl} = \underset{(1.36)}{2.05} - \underset{(.0102)}{.0067}\, atndrte - \underset{(0.48)}{1.63}\, priGPA - \underset{(.098)}{.128}\, ACT + \underset{(.101)}{.296}\, priGPA^2 + \underset{(.0022)}{.0045}\, ACT^2 + \underset{(.0043)}{.0056}\, priGPA{\cdot}atndrte$$
$$n = 680, R^2 = .229, \bar{R}^2 = .222.$$

$atndrte$'nin katsayısı (-0.0067), denklemde interaction terimi olduğu için $priGPA = 0$ iken geçerli olan etkiyi ölçüyor. $priGPA$ serisinde ise sıfır bulunmamaktadır, dolayısıyla, β_1'in negatif işaretli olması önemli değildir. Bu katsayı tek başına devamlılığın etkisini ölçmüyor.

Etkileşim Terimli Modeller: Örnek, attend.gdt

$$\widehat{stndfnl} = \underset{(1.36)}{2.05} - \underset{(.0102)}{.0067}\, atndrte - \underset{(0.48)}{1.63}\, priGPA - \underset{(.098)}{.128}\, ACT + \underset{(.101)}{.296}\, priGPA^2 + \underset{(.0022)}{.0045}\, ACT^2 + \underset{(.0043)}{.0056}\, priGPA{\cdot}atndrte$$
$$n = 680, R^2 = .229, \bar{R}^2 = .222.$$

- Derse devamın finale etkisini β_6 veriyor. β_1 ve β_6 tek tek t testini geçemedikleri halde ikisinin aynı anda sıfır olduğu testi ($H_0 : \beta_1 = \beta_2 = 0$, F testi) reddedilmektedir.
- $priGPA$'nın ortalaması 2.59'dur. Bunu kısmi etkide yerine yazarsak:

$$\Delta stndfnl = -0.0067 + (0.0056)(2.59) = 0.0078$$

- Yorum: $priGPA$ 2.59 iken, devam oranında yüzde 10 puanlık bir artış final notlarında 0.078 standart sapma kadar bir artış yaratır.

Etkileşim Terimli Modeller: Örnek, attend.gdt

- $priGPA$'nin ortalama değerinde devam oranının kısmi etkisi 0.0078 olarak hesaplandı. Bu etki istatistik bakımından sıfırdan farklı mı?
- Bunu test etmek için etkileşim terimini $(priGPA - 2.59)atndrte$ olarak yeniden formüle edip regresyonu yeniden tahmin edeceğiz.
- Bu regresyonda $atndrte$ katsayısı (β_1) $priGPA$ ortalama değerindeki kısmi etkiyi verecektir.
- Standart t testiyle anlamlılığını sınayabiliriz.

Etkileşim Terimli Modeller: Örnek, attend.gdt

$$\widehat{\text{stndfnl}} = \underset{(1.36)}{2.05} + \underset{(0.0026)}{\mathbf{0.0078}}\,\mathbf{atndrte} - \underset{(0.481)}{1.6285}\,\text{priGPA} + \underset{(0.101)}{0.2959}\,\text{priGPA}^2$$
$$- \underset{(0.098)}{0.1280}\,\text{ACT} + \underset{(0.0022)}{0.0045}\,\text{ACT}^2 + \underset{(0.004)}{0.0056}\,\mathbf{(priGPA-2.59)}\cdot\,\mathbf{atndrte}$$
$$n = 680 \quad \bar{R}^2 = 0.2218 \quad F(6, 673) = 33.250 \quad \hat{\sigma} = 0.87287$$

- $atndrte$ katsayısının yorumu: $priGPA$'nin ortalama değerinde (2.59), devam oranındaki yüzde 1 puanlık artış bağımlı değişkende yaklaşık 0.0078 standart sapma artışa yol açmaktadır.
- Devam oranındaki artış yüzde 10 puan ise standardize edilmiş final sınavı sonuçları ortalamasından yaklaşık 0.078 puan kadar uzaklaşır (artar).
- $t = 0.0078/0.0026 = 3$, $H_0 : \beta_1 = 0$ Red, anlamlı $(p - value = 0.003)$.

Uyum Derecesinin Ölçümü: R^2

- R^2, "popülasyonda y'deki değişimin, $x_1, x_2, \ldots, x_k$ tarafından açıklanan yüzdesinin bir tahmini"dir.
- R^2'nin düşük çıkması SEKK (OLS) varsayımlarının ihlal edildiği anlamına gelmez.
- Bağımsız değişken sayısı (k) arttıkça R^2 yükselir. Dolayısıyla, uygun regresyonu seçerken R^2nin kullanımı sınırlı olacaktır.
- Ancak, F testinden hatırlanacağı üzere, yeni bir değişken eklerken R^2'deki görece artış karar kriterimizi oluşturmaktadır.

Düzeltilmiş (Adjusted) R^2

- R^2 aşağıdaki gibi tanımlanıyordu:

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

- Son terimin pay ve paydasını n'e bölersek:

$$R^2 = 1 - \frac{SSR/n}{SST/n} = 1 - \frac{\sigma_u^2}{\sigma_y^2}$$

- R^2 popülasyonda y'deki değişmenin x'lerce açıklanan kısmıdır.
- Ancak SST/n ve SSR/n sapmalı tahminciler olduğundan bunların yerine şu büyüklükleri yazacağız:

$$\frac{SST}{n-1}, \quad \frac{SSR}{n-k-1}$$

Düzeltilmiş (Adjusted) R^2

- Düzeltilmiş R^2

$$\bar{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-k-1}$$

- Düzeltilmiş R^2, ya da R-bar-kare, regresyona yeni bir değişken eklendiğinde artabilir ya da azalabilir. R^2 ise hiç azalmıyordu.
- Bunun nedeni yeni bir değişken eklendiğinde SSR düşerken serbestlik derecesinin de (n-k-1) azalmasıdır.
- Bu yüzden, yeni değişkeni regresyona dahil edip etmemeye karar verirken R^2'yi değil, R-bar kareyi kullanacağız.
- Yeni bir x değişkeni eklendiğinde R-bar kare, ancak ve ancak, bu yeni değişkenin katsayısının t değeri mutlak olarak 1'den büyükse artar.
- Bunu genelleştirirsek, bir gurup x değişkenleri regresyona eklendiğinde, eğer bu yeni değişkenlerin ortak anlamlılık testinde F istatistiği 1'den büyükse R-bar kare artacaktır, aksi halde artmayacaktır.

Örnek

Model 1: OLS, using observations 1–506
Dependent variable: lprice

	Coefficient	Std. Error	t-ratio	p-value
const	8.95348	0.181147	49.4266	0.0000
lnox	−0.304841	0.0821638	−3.7102	0.0002
proptax	−0.00760708	0.000977765	−7.7801	0.0000
rooms	0.288707	0.0181186	15.9343	0.0000

Mean dependent var	9.941057	S.D. dependent var	0.409255
Sum squared resid	36.70511	S.E. of regression	0.270403
R^2	0.566042	Adjusted R^2	0.563449
$F(3, 502)$	218.2650	P-value(F)	1.35e–90

Örnek

Model 2: OLS, using observations 1–506
Dependent variable: lprice

	Coefficient	Std. Error	t-ratio	p-value
const	8.85532	0.172131	51.4452	0.0000
lnox	−0.275421	0.0779513	−3.5332	0.0004
proptax	−0.00422185	0.00102745	−4.1090	0.0000
rooms	0.281587	0.0171939	16.3771	0.0000
crime	−0.0124893	0.00163861	−7.6219	0.0000

Mean dependent var	9.941057	S.D. dependent var	0.409255
Sum squared resid	32.89123	S.E. of regression	0.256225
R^2	0.611133	Adjusted R^2	**0.608028**
$F(4, 501)$	196.8397	P-value(F)	2.7e–101

Örnek

Model 3: OLS, using observations 1–506
Dependent variable: lprice

	Coefficient	Std. Error	t-ratio	p-value
const	9.76749	0.222071	43.9837	0.0000
lnox	−0.355701	0.0763150	−4.6610	0.0000
proptax	−0.00185202	0.00106268	−1.7428	0.0820
rooms	0.251409	0.0172902	14.5405	0.0000
crime	−0.0122323	0.00158140	−7.7351	0.0000
stratio	−0.0370699	0.00599178	−6.1868	0.0000

Mean dependent var	9.941057	S.D. dependent var	0.409255
Sum squared resid	30.55237	S.E. of regression	0.247194
R^2	0.638785	Adjusted R^2	**0.635173**
$F(5, 500)$	176.8435	P-value(F)	4.2e–108

crime ve **stratio** için dışlama testi: $F = 50.35$, $pval < 0.0001$

Düzeltilmiş R^2

- İki modelin R^2'leri karşılaştırılırken bağımlı değişkenlerin aynı olması gerekir.
- Birbirlerinin özel hali olmayan (yuvalanmamış) modellerin seçiminde düzeltilmiş R^2 kullanılabilir (bağımlı değişken aynı ise).
- Örneğin şu iki modeli düşünelim:

$$y = \beta_0 + \beta_1 \log(x), \quad \bar{R}^2_A$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2, \quad \bar{R}^2_B$$

- A modeli daha az parametre içermektedir. Düzeltilmiş determinasyon katsayısı büyük olan model tercih edilebilir.
- B modelinde bağımlı değişken farklı olsaydı, (örneğin $\log(y)$) model seçiminde R^2 ya da $\bar{R}^2$ kullanılamazdı.

Çok Fazla Faktörün Kontrol Edilmesi

- Determinasyon katsayısını yükseltmek adına modele çok sayıda değişken eklemek isteyebiliriz. Ancak bunu yaparken teorik modeli ve ceteris paribus yorumunu göz ardı etmemeliyiz.
- Örneğin, alkol kullanımı ile trafik kazaları arasındaki ilişkiyi çalıştığımızı düşünelim. Daha spesifik olursak alkol satış vergisi ile ölümlü trafik kazaları arasındaki ilişkiyi belirlemek istiyoruz.
- Bağımlı değişken: fatalities (trafik kazalarında ölen kişi sayısı) açıklayıcı değişken: tax (alkollü içecek satış vergisi)
- Bu regresyonu kurarken açıklayıcı değişken olarak alkol tüketimini (beercons) kullanmamamız gerekir. Eğer beercons regresyona tax ile birlikte konursa tax katsayısı şunu verir: bira tüketimi aynı olan iki bölgede vergi oranı yüzde 1 puan arttırılırsa trafik kazalarında ölüm sayısı ne kadar azalır?
- Ancak bizi ilgilendiren bu değildir.

Çok Fazla Faktörün Kontrol Edilmesi

- Bizi ilgilendiren alkollü içecek vergisindeki yüzde 1 puan artışın trafik kazalarını ne kadar azalttığıdır.
- Bunun için regresyona bölgelerin demografik özelliklerini, bireylerin zevk ve tercihlerini vs. yansıtan değişkenler eklenebilir.
- Örneğin erkek nüfusun oranı, 16-21 yaşındakilerin oranı, vs.
- Başka bir örnek: okul kalitesi (Scqual) ve alınan eğitim süresi (EDUC) aynı regresyona açıklayıcı değişken olarak giriyor. Eğer, iyi okul kalitesi zaten zorunlu olarak eğitim süresinin uzun olmasını gerektiriyorsa ve aralarında böyle bir ilişki varsa bizim EDUC değişkenini regresyona sokmamamız lazım.

Yeni Değişkenlerin Eklenmesi

- Yeni bir açıklayıcı değişken bir yandan hata varyansını, σ_u^2, azaltırken, diğer yandan eğer mevcut x'lerle ilişkili ise çoklu-bağıntıyı artırabilir.
- Bu nedenle yeni değişken eklerken çoklu-bağıntı yaratılıp yaratılmadığına dikkat etmek gerekir.
- y ile ilişkili, dolayısıyla da, σ_u^2'yi düşüren yeni bir değişken eğer diğer x'lerle ilişkisiz ise mutlaka regresyona alınmalıdır.
- Örneğin, bira tüketiminin fiyat esnekliğini tahmin etmek isteyelim.

$$\log(beercons) = \beta_0 + \beta_1 \log(price) + u$$

- Bu regresyona tüketicilerin bireysel karekteristikleriyle ilgili değişkenler eklersek (yaş, eğitim vs.) bu değişkenler hem bira talebiyle ilişkili hem de fiyatla ilişkisiz oldukları için hata terimlerinin varyansını büyük ölçüde düşürebilecektir.

Bağımlı Değişkenin Kestirimi (Prediction)

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

- Yukarıdaki örneklem regresyon fonksiyonunda x'lere belirli değerler verirsek $\mathrm{E}(y|x)$ koşullu beklentisinin bir tahminini (kestirimini/prediction) yapmış oluruz.
- Bu belirli değerleri şu şekilde tanımlayalım: $x_1 = c_1$, $x_2 = c_2, \ldots, x_k = c_k$. y'nin kestirim değerini θ_0 ile gösterelim:

$$\theta_0 = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \ldots + \beta_k c_k$$

$$= \mathrm{E}[y|x_1 = c_1, x_2 = c_2, \ldots, x_k = c_k]$$

- θ_0'ın OLS tahmincisi:

$$\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \ldots + \hat{\beta}_k c_k$$

Bağımlı Değişkenin Ortalamasının Kestirimi (Prediction)

- θ_0 için %95 güven aralığı aşağıdaki gibi tanımlanır:

$$\hat{\theta}_0 \pm 2\ se(\hat{\theta}_0)$$

- Bunu hesaplayabilmemiz için $\hat{\theta}_0$'nın standart hatasını bilmemiz gerekir.
- Bir yardımcı regresyon kurarak bunu kolayca hesaplayabiliriz. Tanım gereği:

$$\beta_0 = \theta_0 - \beta_1 c_1 - \beta_2 c_2 - \ldots - \beta_k c_k$$

- Bunu regresyonda yerine yazıp yeniden düzenlersek

$$y = \theta_0 + \beta_1(x_1 - c_1) + \beta_2(x_2 - c_2) + \ldots + \beta_k(x_k - c_k) + u$$

- Bu regresyonda sabit terimin standart hatası ilgili kestirim değerinin standart hatasını verecektir.

Kestirim: Örnek, gpa2.gdt

$$\widehat{colgpa} = \underset{(0.075)}{1.493} + \underset{(.00007)}{.00149}\ sat - \underset{(.00056)}{.01386}\ hsperc - \underset{(.01650)}{.06088}\ hsize + \underset{(.00227)}{.00546}\ hsize^2$$

$$n = 4{,}137,\ R^2 = .278,\ \bar{R}^2 = .277,\ \hat{\sigma} = .560,$$

- Kestirim noktaları: $sat = 1200, hsperc = 30, hsize = 5$ değerleri için bağımlı değişkenin değerini tahmin edelim ve %95'lik güven aralığı kuralım:(hsrank:rank in class; hsize:size of class; hsper:100*(hsrank/hssize)
- Bu değerleri yerine koyduğumuzda $colGPA = 2.70$ olarak bulunuyor.
- Bunun standart hatasını hesaplamak için şu değişkenleri oluşturuyoruz: $sat0 = sat - 1200$, $hsperc0 = hsperc - 30$, $hsize0 = hsize - 5$, $hsizesq0 = hsize - 25$. Daha sonra colGPA'nin bu değişkenler üzerine regresyonunu kuruyoruz.

Kestirim: Örnek, gpa2.gdt

sat0 = *sat* − 1,200, *hsperc0* = *hsperc* − 30, *hsize0* = *hsize* − 5, and *hsizesq0* = *hsize*^{2} − 25.

$$\widehat{colgpa} = \underset{(0.020)}{2.700} + \underset{(.00007)}{.00149}\ sat0 - \underset{(.00056)}{.01386}\ hsperc0 - \underset{(.01650)}{.06088}\ hsize0 + \underset{(.00227)}{.00546}\ hsizesq0$$

$$n = 4{,}137,\ R^2 = .278,\ \bar{R}^2 = .277,\ \hat{\sigma} = .560.$$

- %95 Güv Aralığı:

$$2.70 \pm 1.96(0.020) = [2.66, 2.74].$$

- $\hat{\theta}_0$'nın varyansı en küçük değerine x değişkenleri kendi ortalamalarına eşitlendiğinde ulaşır.
- Öyleyse c_j'ler x_j'lerin ortalama değerlerine ne kadar uzaksa tahminin standart hatası da o kadar büyük olur.
- Yukarıda hesapladığımız standart hata ve güven aralığı y'nin koşullu ortalaması için oluşturulmuştur.
- **Bireysel** y tahminleri için elde edilen güven aralıkları daha geniş olur.

Bağımlı Değişkenin Bireysel Değerlerinin Kestirimi (Prediction)

- Bireysel y değerlerinin kestiriminde hem $\hat{y}$'daki hem de u'daki değişkenlik (varyans) rol oynamaktadır.
- y^o örneklemde yer almayan yeni bir bireyi (kişi, firma, bölge, ülke vs.) temsil etsin:

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

- y^o'ın belirli x_j^o değerleriyle elde edilen OLS kestirimi:

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

- Kestirim hatası:

$$\hat{e}^o = y^o - \hat{y}^o = \beta_0 + \beta_1 x_1^o + \beta_2 x_2^o + \ldots + \beta_k x_k^o + u^o - \hat{y}^o$$

- Beklenen değeri alınırsa

$$\mathrm{E}(\hat{e}^o) = 0$$

Bağımlı Değişkenin Bireysel Değerlerinin Kestirimi (Prediction)

- Tahmin hatasının varyansı

$$\text{Var}(\hat{e}^o) = \text{Var}(\hat{y}^o) + \text{Var}(u^o) = \text{Var}(\hat{y}^o) + \sigma^2$$

- $\text{Var}(\hat{y}^o)$ örneklem hacmi n ile ters orantılıdır, n arttıkça küçülür.
- σ^2 ise gözlenemeyen hata teriminin varyansıdır ve n arttıkça azalmaz.
- Bu nedenle kestirim hatasının varyansını büyük oranda σ^2 belirler.
- Kestirim hatasının standart hatası:

$$se(\hat{e}^o) = \sqrt{\text{Var}(\hat{y}^o) + \hat{\sigma}^2}$$

- %95 güven aralığı:

$$[\hat{y}^o \pm t_{0.025} \cdot se(\hat{e}^o)]$$

Bağımlı Değişkenin Bireysel Değerlerinin Kestirimi: Örnek

- Bireysel kestirim için güven aralıkları ortalama kestirimi için oluşturduğumuz aralıklardan çok daha geniş olur. Çünkü $\hat{\sigma}^2$, $\text{Var}(\hat{y}^o)$'den çok daha büyüktür.
- Örnek olarak, $sat = 1200, hsperc = 30, hsize = 5$ olan bir lise öğrencisi için gelecekteki üniversite başarısını (colGPA) kestirerek %95 güven aralığını oluşturalım.
- Bu değerleri regresyonda yerine koyduğumuzda $colGPA = 2.70$ $(\hat{y}^o)$ olarak bulunuyor.
- Önceki regresyon sonuçlarına baktığımızda $se(\hat{y}^o) = 0.02$ ve $\hat{\sigma} = 0.56$ olduğunu görüyoruz. Bu değerleri yerine koyarsak $se(\hat{e}^o) = \sqrt{0.02^2 + 0.56^2} = 0.56$ %95 güven aralığı

$$2.70 \pm 1.96 \cdot (0.56) = [1.6,\ 3.8]$$

- Bu oldukça geniş bir güven aralığıdır ve muğlaklık regresyonun varyansının büyüklüğünden kaynaklanmaktadır.