Yıldız Teknik Üniversitesi
İktisat Bölümü
Ekonometri II Ders Notları

Ders Kitabı: J.M. Wooldridge, *Introductory Econometrics A Modern Approach*, 2nd. ed., 2002, Thomson Learning.

Ch. 7:
Çoklu Regresyon Modelinde Nitel Değişkenler

CH 7 : REGRESYON ANALİZİNDE NİTEL BİLGİ
İkili (*binary*) ya da kukla (*dummy*) değişkenler

- Önceki bölümlerde bağımlı ve bağımsız değişkenlerimiz her zaman nicel (*quantitative*) nitelikte bilgiler içeriyorlardı: ücretler, iş deneyimi, ev fiyatları, oda sayısı vs.
- Ancak, uygulamada çoğu kez nitel (qualitative) değişkenleri de regresyona dahil etmek zorundayız.
- Bireylerin cinsiyeti, ırkı, dini, doğduğu bölge (kuzey/güney ya da batı/doğu gibi), eğitimi (lise/yüksek gibi)… Tüm bu değişkenler nitel türde enformasyon içermektedirler.

Bölüm Planı

- Altbölüm 7.1: Nitel Bilginin Betimlenmesi
- Altbölüm 7.2: Tek Kukla Açıklayıcı Değişken
- Altbölüm 7.3: Çok Kategorili Kukla Değişkenler
- Altbölüm 7.4: Etkileşimli Kukla Değişkenler
- Altbölüm 7.5: Bağımlı Değişkenin Nitel Olduğu Durum (Doğrusal Olasılık Modeli)

Nitel Bilgi

- Nitel enformasyon çoğu kez ikili (binary) yapı gösterir :
  - erkek/kadın, arabası olan/olmayan, beyaz/beyaz olmayan, yerli/yabancı, ölüm cezası olan/olmayan ülkeler, vb…
- Bu tür enformasyonu **ikili (binary**) değişkenlerle ifade edeceğiz. Bunlara "**sıfır/bir" (0/1**) değişkenler ya da "**gölge ya da kukla (dummy**)" değişken de denir.

Nitel Bilgi

- Hangi kategoriye 1 ya da 0 dediğimiz regresyon sonuçlarını değiştirmez, ancak yorum yapabilmemiz için hangi kategoriye 1 hangisine 0 dediğimizi bilmemiz gerek.
- Örnek :

Cinsiyet kuklası : kadın=1, erkek=0,

Evlilik kuklası : evli=1, evli değil=0

- Nitel kategorileri ayırt etmek için aslında 1/0 yerine başka iki değer de kullanabilirdik. Ancak 1/0 ile gösterildiklerinde katsayılarının yorumu çok kolaylaşmaktadır.

Örnek Veri Seti: cinsiyet ve medeni durum bilgisini içeren ücret verileri

Table 7.1

A Partial Listing of the Data in WAGE1.RAW

person	*wage*	*educ*	*exper*	*female*	*married*
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
. . .	. . .	. . .	. . .	. . .	. . .
525	11.56	16	5	0	1
526	3.50	14	5	1	0

TEK KUKLA DEĞİŞKENLİ REGRESYON

- İkili (nitel) enformasyonu regresyona nasıl dahil edeceğiz?
- X'lerden birisi kukla değişken olsun. Örnek:

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u. \qquad (7.1)$$

- Female=1 olan bireyler kadın, female=0 olan bireyler ise erkek olacaktır.
- δ_o= Aynı eğitim durumuna sahip bir erkekle bir kadının saat ücretleri arasındaki fark.
- Eğer δ_o sıfırdan farklı çıkmazsa (t testi yapacağız) erkek-kadın arasında ücret farklılığı yok demektir.

- Tersine, $\delta_o<0$ çıkarsa, kadınlar erkeklere göre daha düşük saat başına ücretle çalışıyor demektir (*discrimination*).Burada kıyaslama *ceteris paribus* (diğer her şey ayni iken) koşulu altında yapılmaktadır.
- "**E(u | female, educ)=0**" olduğu için şöyle yazabiliriz:

δ_o=E(wage |female=1, educ)- E(wage |female=0, educ).

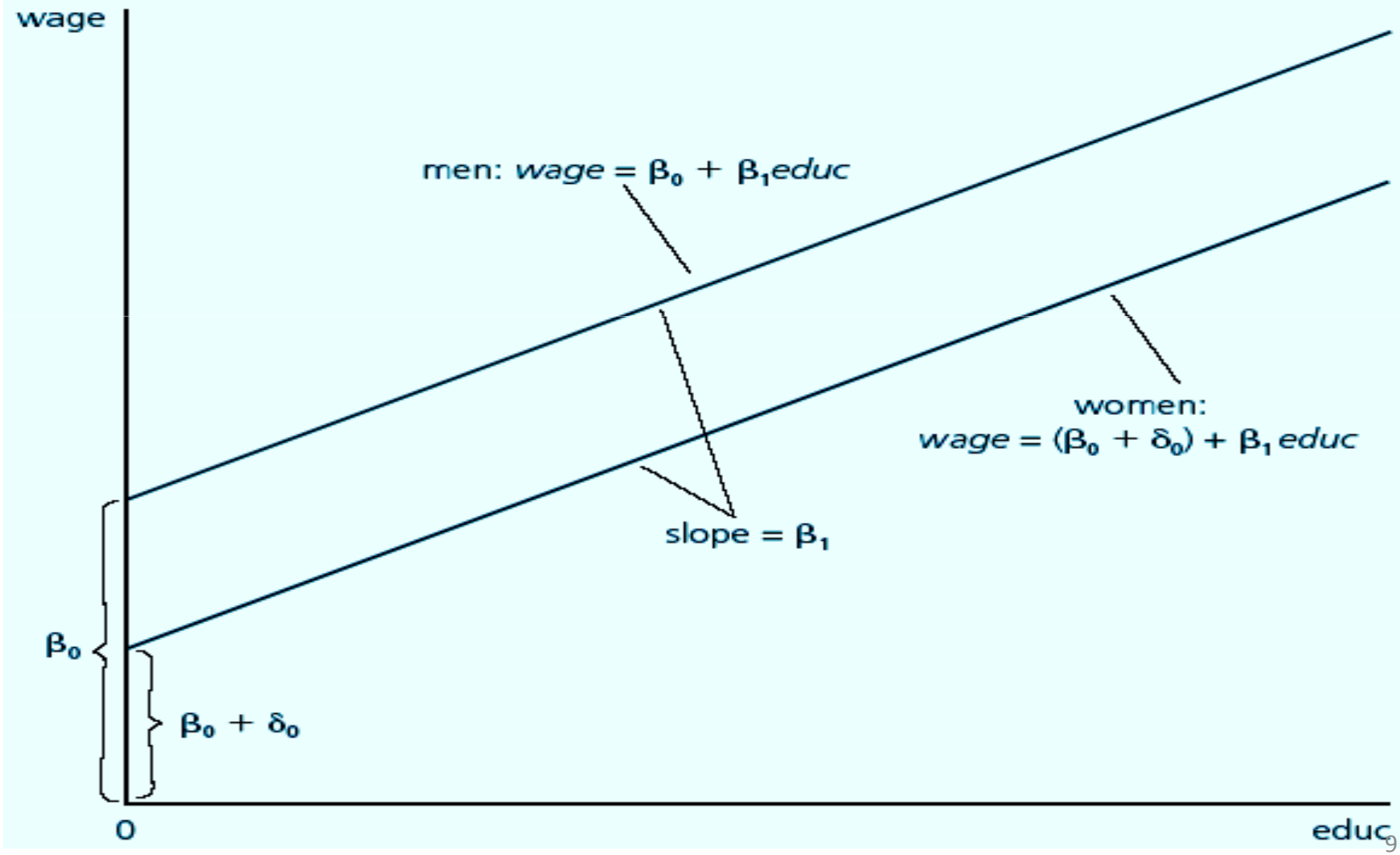
Female=0 → male olduğundan:

$$\delta_0 = \mathrm{E}(wage|female,educ) - \mathrm{E}(wage|male,educ). \qquad (7.2)$$

Sonucu, regresyonun sabit teriminde bir kayma (intercept shift) olarak görürüz: $\delta_o<0$

Figure 7.1

Graph of $wage = \beta_0 + \delta_0 female + \beta_1 educ$ for $\delta_0 < 0$.



- Bir kukla değişken (dummy) iki kategoriyi birbirinden ayırabilmektedir.
- Dolayısıyla, erkek için ayrı, kadın için ayrı bir kukla oluşturmuyoruz. Tek bir kukla ile iki kategoriyi ayırıyoruz.
- "Gerekli kukla sayısı=kategori sayısı-1" formülünü kullanacağız.
- Grafik 7.1 de erkek için sabit terim β_o, kadın için ise $\beta_o+\delta_o$'dır.Kaç kategori (grup) varsa o kadar da sabit (intercept) olacaktır.

- Erkek/kadın için 2 dummy kullansaydık tam çoklu bağıntı (perfect multicollinearity) oluşacaktı ve regresyonu tahmin edemeyecektik. Zira "*female+male=1*" dir.
- Buna kukla değişken tuzağı (*dummy variable trap)* denir.
- Sıfır değerini alan kategori (7.1 de male) baz ya da kıyaslama (base or benchmark) kategorisidir.
- β_o, baz kategorisinin sabitidir.
- δ_o ise, baz kategorisi ile diğer kategorinin sabitleri arasındaki farktır.

- Kadınların baz kategori olmasını istiyorsak modeli şöyle yazarız :

$$wage = \alpha_0 + \gamma_0 male + \beta_1 educ + u,$$

where the intercept for females is α_0 and the intercept for males is $\alpha_0 + \gamma_0$; this implies that $\alpha_0 = \beta_0 + \delta_0$ and $\alpha_0 + \gamma_0 = \beta_0$. In any application, it does not matter how we choose the base group, but it is important to keep track of which group is the base group.

- Diğer bir alternatif, regresyonda hiç sabit (intercept) kullanmayıp her bir kategori için bir dummy kullanmaktır.

$$\text{wage} = \delta_o \text{ female} + \gamma_0 \text{ male} + \beta_1 \text{ educ} + u$$

- Burada female = 1 eğer çalışan kadınsa, =0 değilse
- male = 1 eğer çalışan kadınsa, =0 değilse
- Bu durumda kukla değişken tuzağına düşmeyiz.
- Çünkü her bir kategori için bir sabit (intercept) tahmin etmiş oluruz.
- Ancak sabitsiz bir regresyonda R^2 hesaplanamamaktadır.
- Üstelik iki sabitin birbirinden farklı olup olmadığının testi de zordur. Dolayısıyla, bu alternatife pek başvurmayacağız.

- Birden fazla açıklayıcı değişken olması durumu değiştirmez. Örneğin;

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u. \qquad (7.3)$$

Eğer "Ho: δ_o=0" boş hipotezi "H_1: δ_o<0" alternatif hipotezine karşı testi geçemezse, kadınlara karşı ücret ayrımcılığı (*discrimination*) var demektir.

- δ_o için t testi yaparak anlamlı olup olmadığına karar vereceğiz. Bu bakımdan gölge değişkenlerin diğer değişkenlerden bir farkı yoktur.
- Fark sadece gölge değişkenlerin katsayılarının yorumundadır.

- **Örnek 7.1**

$$\widehat{wage} = \underset{(0.72)}{-1.57} - \underset{(0.26)}{1.81}\, female + \underset{(.049)}{.572}\, educ + \underset{(.012)}{.025}\, exper + \underset{(.021)}{.141}\, tenure \quad (7.4)$$

$$n = 526,\ R^2 = .364.$$

- Burada, *female* kukla değişkeninin katsayısı **-1.81**, eğitim, tecrübe ve son işyerindeki kıdemi aynı olan bir kadınla bir erkeğin saat ücretleri arasındaki ortalama farkı ölçer. Bu fark saat ücretlerde 1.81$ kadardır.
- Regresyon sabiti **(-1.57**), female=0 iken geçerli sabit, yani erkeklere ait regresyonun sabitidir. Educ=exper=tenure=0 iken ücretin ne olduğunu gösterdiği için bu örnekte fazla bir anlam ifade etmemektedir.

- (7.4) de *female* değişkeninin katsayısının ücret farkını gösterdiğini şöyle de kanıtlayabiliriz:
- (7.4) de tüm açıklayıcı değişkenleri dışarıda bırakıp sadece female'i tutalım. Regresyon şöyle tahmin edilmektedir.

$$\widehat{wage} = \underset{(0.21)}{7.10} - \underset{(0.30)}{2.51}\ female \qquad (7.5)$$
$$n = 526,\ R^2 = .116.$$

- Burada sabit (intercept), 7.10, female=0 iken geçerli olan sabittir, yani erkeklere ait ortalama saat-ücrettir. Kadınlara ait ortalama saat-ücret ise:

7.10 - 2.51(1) = 4.59 $

olur.

- Fark **7.10 - 4.59 = 2.51$** = β(female)

- Yukarıdaki örneğe 274 erkek, 252 kadın dahil edilmiştir. Dolayısıyla, kategoriler dengelidir (*balanced*).
- (7.5) deki regresyon kategorilere ait ortalamaların birbirinden farklı olup olmadığını test etmenin basit bir yoludur.

- Ortalama kadın ve erkek saat-ücretleri farkı -2.51$ dır ve standart hatası 0.30 olan bu katsayının t değeri -8.37 (=-2.51/0.3) gibi çok yüksek (mutlak olarak) bir değerdir.
- Yani, istatistiksel olarak %1 düzeyinde bile anlamlıdır.

- Bu t testinin geçerliliği sabit varyans (homoscedasticity) varsayımının sağlanmasına bağlıdır.
- Yani,erkeklerin ücretlerine ait varyans (populasyonda) kadın ücretleri varyansı ile aynı olmalıdır ki test yapabilelim.
- Cinsiyetler-arası ücret farkı (7.4) de (7.5)dekinden daha az çıkmıştır. Çünkü, ilk regresyonda bireylere ait diğer özellikler (educ, exper , tenure) regresyona dahil edilerek kontrol altına alınmıştır.
- Eğitim, deneyim gibi faktörler (ki bu değişkenlerin ortalaması kadınlarda daha düşüktür) dikkate alındığında ücret farkı biraz daha azalmaktadır.

- Kukla değişkenler her zaman cinsiyet gibi önceden belirlenmiş (*predetermined*) kategorileri değil, bireylerin seçimleri, mülkiyet durumu vb. farklılıkları da temsil ederler.
- Bu durumlarda **nedensellik (*causality*)** temel konudur.
- Örneğin, kişisel bilgisayarı (PC) olan üniversite öğrencilerinin GPA puanlarının bilgisayarı olmayanların puanlarından daha yüksek olduğunu saptamış olalım.
- Bu neyi gösterir? Nedenselliğin yönü nedir? "Bilgisayar sahipliği→ yüksek GPA" mi yoksa "Yüksek GPA→ bilgisayar sahipliği" mi?

- Demek ki, nedenselliğin yönünü tayin edebilmek için öğrencilerin başarı durumları ile ilgili bazı değişkenleri kontrol altına almamız gerektir. Bunlar lise bitirme notları, genel yetenek testi sonuçları vs. olabilir.
- Bu değişkenleri regresyona dahil ettiğimizde, bilgisayar sahipliğinin üniversite notuna hala pozitif katkısı çıkıyorsa, nedenselliğin yönü "bilgisayar→ başarı" dır diyebileceğiz.

$$\widehat{colGPA} = \underset{(0.33)}{1.26} + \underset{(.057)}{.157}\,PC + \underset{(.094)}{.447}\,hsGPA + \underset{(.0105)}{.0087}\,ACT \qquad \textbf{(7.6)}$$

$$n = 141,\ R^2 = .219.$$

- hsGPA: lise (high school) not ortalaması, PC=1 bilgisayarı varsa, ACT=yetenek sınavı sonucu

- Regresyonda, lise notu üniversite notunun belirlenmesinde kuvvetle etkili olduğu halde ACT yetenek sınav sonuçları etkili değildir.
- Bu iki değişkenin üniversite notlarına etkisi dikkate alındığında da PC ‘nin katkısı hala pozitiftir ve istatistiksel olarak anlamlıdır.
- Yani, sadece zaten iyi öğrenci (lisede) olanlar PC almamış, bazı öğrenciler PC aldıkları için üniversitede iyi öğrenci haline gelmiştir.
- hsGPA ve ACT değişkenlerini regresyon dışında bıraktığımızda PC’nin katkısı 0.170 (se:0.063) ‘ye çıkıyor.

Kukla Değişkenlerle Etkileşim

- **Farklı eğimler (slopes) elde etmek**
- Yukarıda kukla değişkenler kullanarak aynı regresyonda çok sayıda farklı sabite (intercept) yer verebildik.
- Şimdi farklı eğimler türetmeyi görelim.
- Örneğin, ücret regresyonunda bir yıllık ilave bir eğitimin erkek ve kadın ücretlerinde yaratacağı etkinin farklı olup olmadığını ölçmek isteyelim.
- Bunun için kadın gölge değişkeni *female*'i *educ* değişkeni ile etkileşime (*interaction*) sokacağız.

- Regresyonu şöyle yazacağız:

$$\log(wage) = (\beta_0 + \delta_0 female) + (\beta_1 + \delta_1 female)educ + u. \qquad (7.16)$$

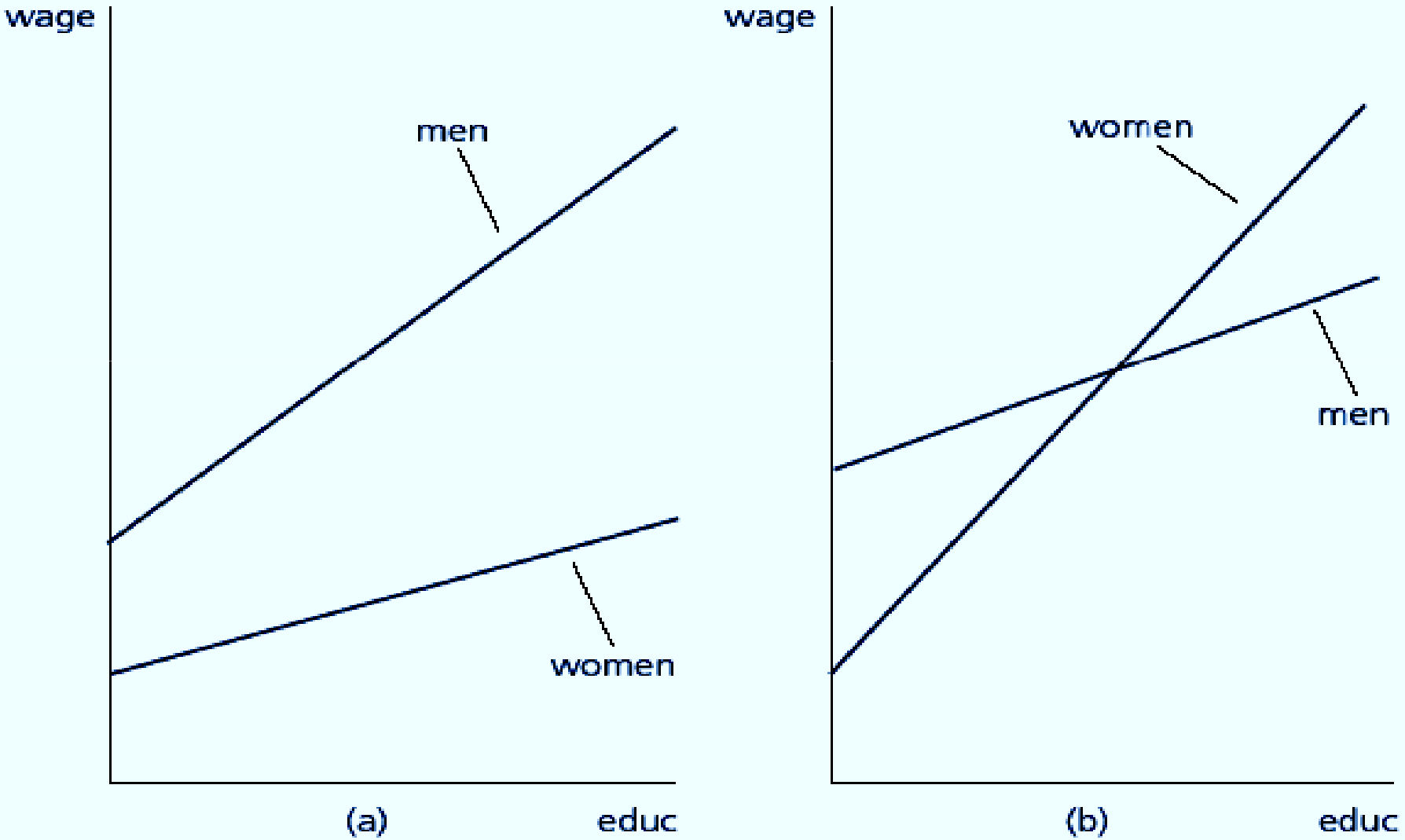
- Bu regresyonda "female = 0" değerini yerine koyarsak, β_0' ın erkekler için sabit, β_1' in ise erkekler için eğim (slope) katsayıları olduğunu görürüz.
- Female=1 aldığımızda (yani, kadın kategorisini seçtiğimizde), "$\beta_0 + \delta_0$", kadın için sabiti,

  "$\beta_1 + \delta_1$" ise kadın için eğim katsayısını gösterecektir.

- Böylece δ_o, iki kategorinin sabitleri arasındaki farkı, δ_1 ise, eğimleri arasındaki farkı verecektir.
- Kadın ve erkekler arasında bir yıllık ilave bir eğitimin ücret getirisi δ_1 kadar farklıdır.
- Regresyonu fark katsayılarının almaşık pozitif ve negatif değerleri için şöyle gösterebiliriz:

Figure 7.2

Graphs of equation (7.16). (a) $\delta_0 < 0$, $\delta_1 < 0$; (b) $\delta_0 < 0$, $\delta_1 > 0$.



- Grafik (a) da, kadınlara ait regresyonun erkeklere ait regresyonunkilere kıyasla hem sabiti, hem de eğimi daha düşüktür.
- Yani, kadınların saat-başına ücretleri tüm eğitim düzeylerinde erkeklerinkinden daha düşüktür.
- Fark (gap) *educ* yükseldikçe artmaktadır.
- Grafik (b) de ise, erkeklere kıyasla, kadın regresyonunun sabiti daha düşük, ancak educ değişkeninin eğimi daha yüksektir. Bu şu anlama gelir: düşük eğitim düzeylerinde kadınlar, yüksek eğitim düzeylerinde ise erkekler daha az kazanmaktadır.

• (7.16) daki modeli female ve educ değişkenleri arasında karşılıklı etkileşim (interaction) kurarak tahmin edeceğiz :

$$\log(wage) = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female \cdot educ + u. \qquad (7.17)$$

• Etkileşim değişkeni **"female*educ"** , erkekler için **0**, kadınlar için ise **educ** değişkeni değerini alacaktır.

• Şu hipotezi test edeceğiz : Eğitimin getirisi erkek ve kadınlar için aynı mı? H_o: $\delta_1 = 0$, H_1: $\delta_1 \neq 0$.

• H_o hipotezi, "log(wage)'in educ'ya göre türevi erkek ve kadınlar için aynıdır" anlamına gelmektedir.

• Burada, sabitlerle ilgili bir kısıt yoktur. Yani, eğitimin getirisinin iki grup için de eşit olduğunu söylüyoruz.

• Eğitim dışındaki faktörlerden dolayı eşitsizlik olabilir.

- Şu hipotezle de ilgiliyiz :

  “Aynı eğitim düzeyine sahip erkek ve kadınların ortalama ücreti aynıdır (Ho)/ farklıdır (H1)”.

$$H_0: \delta_0 = 0, \delta_1 = 0.$$

- Bu hipotez, δ_o ve δ_1 katsaylarının aynı anda sıfıra eşit olup olmadıklarının testidir.

EXAMPLE 7.10
(Log Hourly Wage Equation)

We add quadratics in experience and tenure to (7.17):

$$\widehat{\log(wage)} = \underset{(.119)}{.389} - \underset{(.168)}{.227}\,female + \underset{(.008)}{.082}\,educ - \underset{(.0131)}{.0056}\,female{\cdot}educ + \underset{(.005)}{.029}\,exper - \underset{(.00011)}{.00058}\,exper^2 + \underset{(.007)}{.032}\,tenure - \underset{(.00024)}{.00059}\,tenure^2 \qquad (7.18)$$

$$n = 526,\ R^2 = .441.$$

The estimated return to education for men in this equation is .082, or 8.2%. For women, it is .082 − .0056 = .0764, or about 7.6%. The difference, −.56%, or just over one-half

- Bir yıllık ilave eğitimin getirisi erkek ve kadınlar için aynıdır. "**Female*educ**" etkileşim değişkeninin katsayısının (δ1) t istatistiği -0.0056/0.0131 = -0.43 dür ve istatistiksel olarak anlamsızdır.
- Regresyonda interaction değişkeni bulunduğundan, female değişkeninin katsayısı (δo), educ=0 iken erkek ve kadınlar arasındaki ücret farkını ölçecektir. Örnekte eğitim yılı sıfır olan kişi bulunmadığı için ve ayrıca ***female*** ile ***female*educ*** değişkenleri arasında çoklu-bağıntı (multicollinearity) olduğundan δo'ın standart hatası yüksek, dolayısıyla da t değeri düşüktür (-1.35).

- Bu nedenle, female'in katsayısını şöyle tahmin edeceğiz : Etkileşim değişkeninde ***educ*** yerine onun ortalamasından sapmasını (***educ- ortalama***) kullanacağız. Educ değişkeninin ortalaması 12.5 yıldır.
- Yeni etkileşim değişkenimiz ***"female*(educ-12.5***)" olacaktır. Bu değişiklikten sonra regresyonu yeniden tahmin edeceğiz. Bu durumda, δo, educ=12.5 iken geçerli olan ücret farkını gösterecektir. Yani ortalama eğitim düzeyinde ücret farkı olup olmadığını göreceğiz.
- F testi sonucu, δo ve δ1 ' in ikisinin birden sıfıra eşit olmadığını gösteriyor. O halde sabit terim erkek ve kadın için farklıdır. Dolayısıyla, yukarıdaki (7.9) modelini tercih edeceğiz. İki farklı sabit tahmin ediyordu.

- ÖRNEK : Oyuncunun ırkının ve kentin ırk bileşiminin beyzbol oyuncuları maaşlarına etkisi

- Örnek 330 oyuncu kapsıyor. ***Black*** ve ***hispan***, siyah ve ispanyol kökenli oyuncuları gösteren kukla değişkenleridir. Baz kategori beyaz oyunculardır. ***Percblck*** ve ***perchisp***, takımların ait olduğu kentlerde siyah ve ispanyol kökenli vatandaşların oranıdır.
- Önce dört ırk değişkeninin (*black*, *hispan*, *black*percblck ve hispan*perchisp*) ortaklaşa (jointly) anlamlı olup olmadığına bakalım. F testini , kısıtsız ve kısıtlı modelin R2'lerini kullanarak yapalım.
- 4 değişken dışarıda bırakıldığında R2=0.626 bulunuyor. Kısıt sayısı=4, kısıtsız modelde df=330-13. F istatistiği 2.63 olarak elde edilir (p değeri:0.034). 4 değişken birlikte anlamlıdır.

- Örnek :

$$\begin{aligned}\widehat{\log(salary)} = {} & 10.34 + .0673\,years + .0089\,gamesyr \\ & \quad (2.18) \quad (.0129) \qquad (.0034) \\ & + .00095\,bavg + .0146\,hrunsyr + .0045\,rbisyr \\ & \quad (.00151) \qquad (.0164) \qquad (.0076) \\ & + .0072\,runsyr + .0011\,fldperc + .0075\,allstar \\ & \quad (.0046) \qquad (.0021) \qquad (.0029) \\ & - .198\,black - .190\,hispan + .0125\,black{\cdot}percblck \\ & \quad (.125) \qquad (.153) \qquad (.0050) \\ & + .0201\,hispan{\cdot}perchisp, \quad n = 330,\ R^2 = .638. \\ & \quad (.0098)\end{aligned} \tag{7.19}$$

- Oyuncuların verimliliği ile ilgili tüm değişkenleri sabit tutarak ırk kuklalarını yorumlayalım.
- Black'in katsayısı, -0.198 : diğer her şey sabitken, hiç siyah nüfusun bulunmadığı (*percblck*=0) bir şehirde oynayan siyah oyuncu beyaz oyunculara kıyasla %19.8 daha az kazanmaktadır. O şehirde siyah nüfus oranı arttıkça siyah oyuncunun kazancı artacaktır (*black*percblck* etkileşim değişkeninin katsayısı + işaretli).
- Örneğin, siyahların %20 olduğu bir şehirde siyah oyuncular beyazlardan (baz kategori) **%5.2** (= -0.198 + 0.0125* 0.20 = 0.052) daha fazla ücret alacaktır.

- Benzer şekilde ispanyol kökenli oyuncular, ispanyol kökenlilerin az oranda bulunduğu şehirlerde beyaz oyunculara kıyasla daha az kazanacaklardır.

- Siyah nüfus sabitken, ispanyol nüfusun %9.45 'i geçtiği şehirlerde ispanyol kökenli beyzbolcular beyaz oyunculardan daha fazla kazanacaktır : -0.190 + 0.0201*perchisp = 0 → perchisp = 9.45.

- Oyuncu kazancı ile oyuncunun oynadığı şehrin nüfus yapısı ilişkili çıkıyor. Ancak, ilişkinin yönünü bu regresyondan bulamayız. Yoğun siyah nüfusun yaşadığı şehirlerde siyah oyuncuların daha çok kazanması, en iyi siyah oyuncuların siyahların ağırlıklı olduğu şehirlerde yaşamak istemelerinden ileri gelebilir. Ya da şehir takımları şehrin etnik yapısına göre oyuncuların ücretlerini farklılaştırıyor olabilir.

İki ayrı kategoriyi aynı regresyonla açıklayabilir miyiz?

- İki grup arasında regresyonun tüm katsayılarının farklı olup olmadığını test edelim.
- Üniversitenin erkek ve kız öğrencilerden oluşan atletizm takımı bulunuyor. Soru şu : Bu atletlerin notlarını aynı regresyonla açıklayabilir miyiz?
- Regresyon :

$$cumgpa = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + u,$$

- SAT, genel yetenek testi sonuçları; *hsperc*, lise (high school) mezuniyet sıralaması (%); *tothrs*, alınan toplam ders saati.
- Hem sabit terimin hem de eğimlerin kız ve erkek atletler arasında farklı çıkacağını düşünüyorsak modeli şu hale getiririz :

- Modele sabit ve eğim kuklaları koyuyoruz :

$$cumgpa = \beta_0 + \delta_0 female + \beta_1 sat + \delta_1 female \cdot sat + \beta_2 hsperc + \delta_2 female \cdot hsperc + \beta_3 tothrs + \delta_3 female \cdot tothrs + u. \quad \textbf{(7.20)}$$

- "Erkek ve kız atletlerin notlarını aynı regresyonla açıklayabiliriz" şeklindeki H_o hipotezi şöyle olacaktır :

$$H_0: \delta_0 = 0, \delta_1 = 0, \delta_2 = 0, \delta_3 = 0. \quad \textbf{(7.21)}$$

- Eğer H_1 kabul edilirse, yani katsayıların en az birisi sıfırdan farklı çıkarsa modelin iki grup için ayrı ayrı olacağına karar vereceğiz.

- Tahmin edilen regresyon :

Using the spring semester data from the file GPA3.RAW, the full model is estimated as

$$\widehat{cumgpa} = \underset{(0.21)}{1.48} - \underset{(.411)}{.353}\, female + \underset{(.0002)}{.0011}\, sat + \underset{(.00039)}{.00075}\, female \cdot sat$$
$$- \underset{(.0014)}{.0085}\, hsperc - \underset{(.00316)}{.00055}\, female \cdot hsperc + \underset{(.0009)}{.0023}\, tothrs$$
$$- \underset{(.00163)}{.00012}\, female \cdot tothrs \qquad (7.22)$$
$$n = 366,\ R^2 = .406,\ \bar{R}^2 = .394.$$

- Kısıtlı modelde R2=0.352 bulunuyor. F=8.14 hesaplıyoruz. Bu ise çok yüksek bir değer ve p-değeri sıfıra eşit, Yani, katsayılar aynı anda sıfıra eşit değildir. Erkek ve kadın atletlerin notları aynı model tarafından açıklanamaz. Farklı modeller gerektir.
- Katsayıların tek tek t değerleri 2 ‘den küçük olmakla beraber ortak anlamlılık testini kuvvetle geçebilmektedirler.
- Female’in katsayısını yorumlarken, bu katsayının ***sat=hsperc=tothrs=0*** iken not farkını gösterdiğini unutmayalım. Her üç değişkenin sıfır alınması ilginç bir senaryo olmadığı için regresyonda bu değişkenlerin ortalama değerleri konarak female değişkeninin katsayısı yeniden hesaplanır :

- Sat =1,100, hsperc=10 ve tothrs=50 koyduğumuzda puan farkı: {-0.353+0.00075*1100-0.00055*10-0.00012*50} = **0.461**

  olarak bulunur.

- Yani, kadın atletlerin GPA'leri aynı özelliklere sahip erkeklerinkine kıyasla yarım puana yakın daha yüksektir.

Chow testi

- **Chow İstatistiği**
- Değişken sayısı fazla ise yukarıdaki gibi her bir değişken için interaction yaratmak pratik olmayacaktır. İki grubun aynı modelle açıklanıp açıklanamayacağını Chow testi ile daha kısa yoldan belirleyebiliriz.
- Grup 1=g1, grup 2 = g2 diyelim.

$$y = \beta_{g,0} + \beta_{g,1}x_1 + \beta_{g,2}x_2 + \ldots + \beta_{g,k}x_k + u, \qquad (7.23)$$

- Soru, tüm beta katsayılarının her iki grup için eşit olup olmadıklarıdır.

- Sabit ve eğim farkı gölge değişkenleri SAT dışında anlamlı çıkmamıştır.
- Ancak, yukarıda belirtildiği gibi, biz tek tek t testlerine değil ortak (jointly) anlamlılık testine (F testi) dayanarak karar vereceğiz.
- Tüm sabit ve eğim fark katsayılarının aynı anda sıfır olduğu şeklindeki Ho hipotezini test edebilmek için söz konusu değişkenleri (female ve tüm interaction değişkenleri) dışarıda bırakarak kısıtlı modeli tahmin ediyor ve R2 ya da SSR 'lerden hareketle F istatistiğini hesaplıyorduk.

$$H_0: \delta_0 = 0, \delta_1 = 0, \delta_2 = 0, \delta_3 = 0. \qquad \textbf{(7.21)}$$

Chow istatistiği ve testi

- Değişken sayısı fazla ise yukarıdaki gibi her bir değişken için interaction yaratmak pratik olmayacaktır. İki grubun aynı modelle açıklanıp açıklanamayacağını Chow testi ile daha kısa yoldan belirleyebiliriz.
- Grup 1=g1, grup 2 = g2 diyelim.Modelimiz :

$$y = \beta_{g,0} + \beta_{g,1}x_1 + \beta_{g,2}x_2 + \ldots + \beta_{g,k}x_k + u, \qquad \textbf{(7.23)}$$

- Soru, tüm beta katsayılarının her iki grup için eşit olup olmadıklarıdır.

- Test için şu adımlar izlenir :
- (7.23) ü her iki grup verileri ile ayrı ayrı tahmin ediniz ve elde ettiğiniz regresyonların artık kareler toplamlarını SSR1 ve SSR2 olarak adlandırınız. Bu iki SSR'nin toplamı kısıtsız modelin SSR'sini verecektir : SSRur = SSR1 + SSR2.
- İki grubun verilerini birleştirerek (7.23) ü tek bir regresyon olarak tahmin ediniz ve artık kareler toplamını SSR olarak adlandırınız. SSR, kısıtlı modelin artık kareler toplamıdır : SSR = SSRr.
- Kısıt sayısı k+1 dir (= (7.23) deki parametre sayısı). Kısıtsız modelin serbestlik derecesi, df = sabit kuklası + k değişkenin her birisi için interaction kuklası + regresyonun kendi sabiti + k tane değişken = **n – 2(k+1**) dir.

- Şu formülden Chow istatistiğini hesaplayınız :

$$F = \frac{[SSR - (SSR_1 + SSR_2)]}{SSR_1 + SSR_2} \cdot \frac{[n - 2(k+1)]}{k+1} \qquad (7.24)$$

- Bulunan Chow istatistiğini F tablosundan F {k+1 ; (n-2(k+1)} kritik değeri ile karşılaştırarak karar veriniz : cal F > tab F ise Ho red, aksi halde kabul.
- Ho : βo(g1) = βo(g2), β1(g1)= β1(g2), ... βk(g1)= βk(g2),

- Yukarıdaki GPA örneğine Chow testi uygulayalım :
- Kız atlet sayısı n1= 90, erkek atlet sayısı n2= 276, toplam örnek n =n1+n2 = 366.
- İki grubun verileri bir arada ele alınarak birleştirilmiş regresyon (pooled regression) uygulandığında artık kareler toplamını 85.515 olarak buluyoruz : SSRr (= formüldeki SSR) = 85.515. Kadın atlet regresyonundan SSR1=19.603, erkek regresyonundan SSR2=58.752 elde ediyoruz. Bölece SSRur = 19.603 + 58.752 = 78.355 bulunur.

- F =[(85.515-78.355) / 78.355] [358 / 4] = 8.18 . Yukarıdaki ile aynı sonuca ulaştık. %5 için tab F(4 ; 358) = 2.372 < cal F → Ho red.

- Chow testinin önemli bir yetersizliği, Ho hipotezinde grup modelleri arasında hiçbir farkın olmadığının kabul edilmesidir. Oysa, bazen “eğim katsayıları aynı ancak sabit terim farklıdır” diye test yapmak isteriz.

- Bunu Chow testi ile yapamayız. **Example 7.10** da gördüğümüz gibi regresyona *interaction* terimleri koyup sonra da bu interaction katsayılarının ortak anlamlılık (*joint significance*) F testini yapmamız gerekmektedir. Sabit üzerine kısıt konulmayacaktır.

- GPA örneğinde, **Ho : δ1=δ2=δ3 = 0** için F=1.53 bulunuyor. Bu da ancak p=0.205 düzeyinde anlamlıdır. Yani, Ho ‘ı reddedemiyoruz. O halde, interaction ‘li model i değil sadece sabit kuklasının bulunduğu şu modeli tercih edeceğiz :

$$\widehat{cumgpa} = \underset{(0.18)}{1.39} + \underset{(.059)}{.310}\, female + \underset{(.0002)}{.0012}\, sat - \underset{(.0012)}{.0084}\, hsperc + \underset{(.0007)}{.0025}\, tothrs \qquad (7.25)$$

$$n = 366,\ R^2 = .398,\ \bar{R}^2 = .392.$$

- (7.25) de eğim katsayıları (7.22) de baz grup (erkekler) için bulunan katsayılara yakındır. Interaction terimlerini bırakmamız fazla bir değişmeye yol açmadı.
- Ancak, (7.25) de female'in katsayısı yüksek bir t değerine sahiptir. Yani, aynı SAT, hsperc ve tothrs düzeyine sahip erkek ve kadın atletlerin GPA notları kadınlar lehine 0.31 puan daha yüksektir.

İkili (binary) bağımlı değişken: Doğrusal Olasılık Modeli (The Linear Probability Model)

- Şimdiye kadar gördüğümüz örneklerde bağımlı değişken y hep nicel (quantitative) nitelikte bir değişkendi : ücret, not, Log(ücret) gibi ….
- y, nitel türde ikili (binary) bir değişkense ne olacak ?
- y, 1 ve 0 değerini alan bir değişken olabilir. Yerli/yabancı, evi olanlar/olmayanlar, yabancı dil bilenler/bilmeyenler vb. ikili (binary) türde değişkenler.
- y'nin ikili değişken olduğu şu regresyon neyi ifade edecektir?

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u, \qquad (7.26)$$

- Bu regresyonda y, sadece 1 ve 0 değerlerini aldığı için betaların yorumu eskisi gibi (diğer tüm faktörler sabitken, x(j)'de 1 birimlik değişmenin y'de yol açacağı değişme) olmayacaktır.

- Zira, y sadece 1 den 0'a, 0'dan 1'e değişebilecektir.

- Buna rağmen bu regresyondaki betalar da oldukça yararlı olabilecek yorumlara sahiptirler.

- Eğer sıfır-koşullu ortalama varsayımı, MLR.3,geçerliyse, E (u|x1,…xk)=0, her zamanki gibi şun yazabiliriz :

$$\mathrm{E}(y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

- X, tüm bağımsız değişkenleri gösteren bir vektördür.
- y, 1/0 değerlerini alan ikili bir değişken iken, **P(y=1|x) = E(y|x)** eşitliği bulunmaktadır. Yani, y'nin 1 değerini alması (buna başarı diyelim) olasılığı, y'nin beklenen değeri ile aynı şeydir :

$$\mathrm{P}(y = 1|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \qquad (7.27)$$

- (7.27), başarı olasılığının, buna p(x)=P(y=1|x) diyelim, x'lerin doğrusal bir fonksiyonu olduğunu söyler.

- (7.27), ikili tepki modeline (***binary response model***) bir örnektir. P(y=1|x), tepki olasılığı (*response probability*) olarak da adlandırılır.
- Başarı (y=1) ve başarısızlık (y=0) oranları toplamı 1 olduğu için, başarısızlık olasılığı da x'lerin doğrusal bir fonksiyonudur :

$$P(y = 0|x) = 1 - P(y = 1|x)$$

- y'nin ikili (binary) bir değişken olduğu çoklu-regresyon modeli doğrusal olasılık modeli (l**inear probability model (LPM**) adını alır. Çünkü, P(y=1|x), betaların doğrusal bir fonksiyonudur.

- Bu modelde, β(j)'nin yorumu şöyle olacaktır : Diğer faktörler sabitken, x(j)'de bir birimlik değişme başarı şansında β(j) kadar bir değişme yaratacaktır.

$$\Delta P(y = 1|x) = \beta_j \Delta x_j. \qquad (7.28)$$

- Böylece çeşitli değişkenlerin y=1 olma olasılığına yapacağı katkıyı saptayabileceğiz. SEKK (OLS) nin tüm diğer özellikleri aynen burada da geçerlidir.
- Tahmin edilen regresyon :

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

- Burada yhat, başarı oranı tahminidir. Bo, x'ler sıfırkenki başarı oranı tahminidir.

Örnek: işgücüne katılma [Mroz (1987) verileri]

- Eğer örneğe giren evli kadınlar 1975 yılının herhangi bir ayında işgücüne katılmışsa (ücretli çalışmışsa) inlf =y değişkeni 1, hiç katılmamışsa 0 değerini alsın.
- Diğer değişkenler :

nwifeinc : kocanın geliri (bin$)

kidslt6 : 6 yaşından küçük çocuk sayısı,

kidsge6: 6-18 yaş arasındaki çocuk sayısı

- Regresyon şöyle tahmin ediliyor (753 kadının 428'i işgücüne katılmıştır):

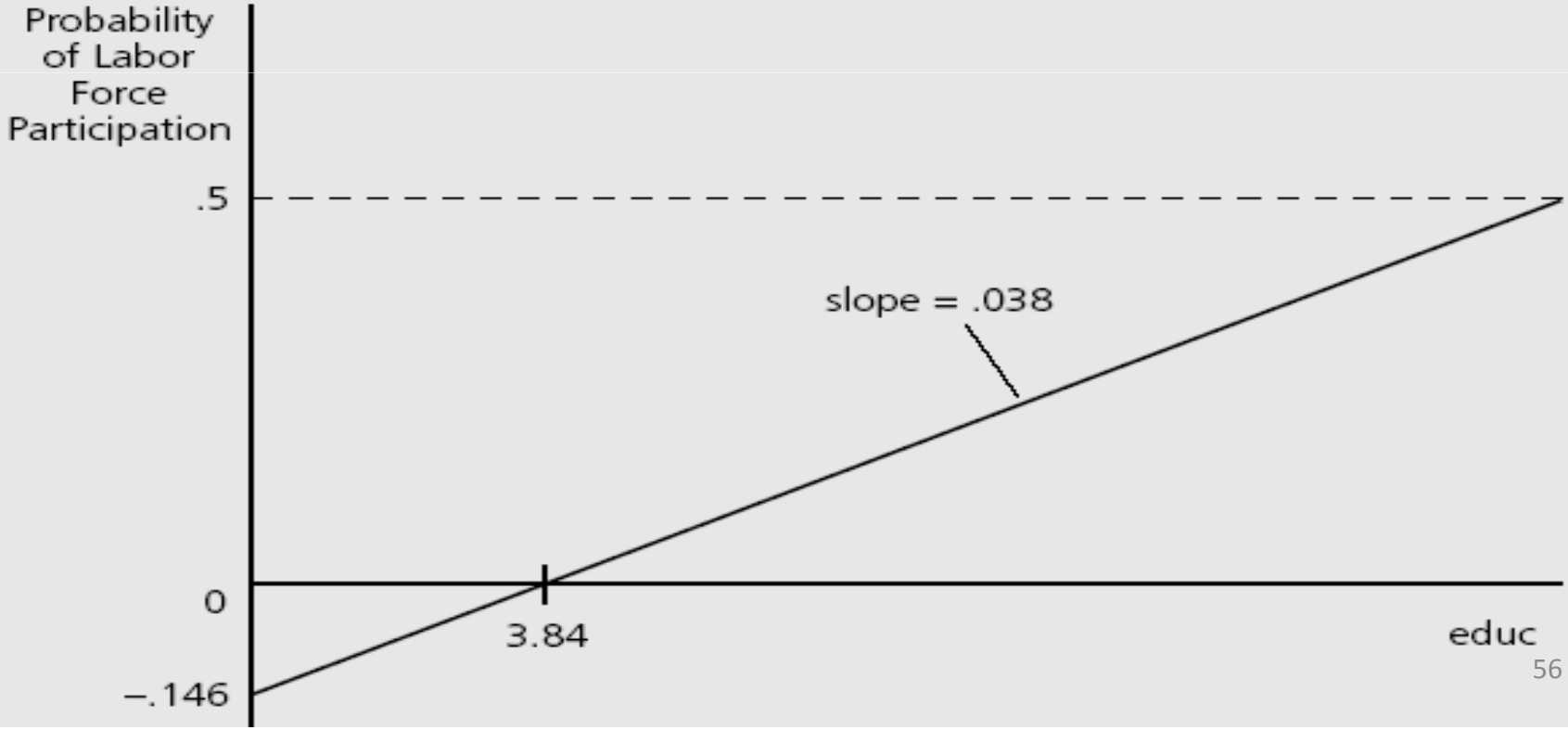
$$\widehat{inlf} = .586 - .0034\ nwifeinc + .038\ educ + .039\ exper$$
$$(.154) \qquad (.0014) \qquad (.007) \qquad (.006)$$
$$- .00060\ exper^2 - .016\ age - .262\ kidslt6 + .0130\ kidsge6 \qquad \textbf{(7.29)}$$
$$(.00018) \qquad (.002) \qquad (.034) \qquad (.0132)$$
$$n = 753,\ R^2 = .264.$$

- *kidsge6* hariç tüm değişkenler istatistiksel olarak anlamlı çıkmıştır ve katsayıların işaretleri iktisat teorisi ya da sağduyu ile uyumludur.
- Betaların yorumları : Örneğin educ değişkeninin betası, 0.038, diğer faktörler sabitken, ilave bir yıl eğitimin işgücüne katılma (inlf=1) olasılığını %3.8 artırdığı anlamına gelmektedir.

- (7.29) da değişkenleri belli değerlerde sabitleyerek (nwifein = 50, exper = 5, age = 30, kidslt6 = 1, kidsge6 = 0), **inlf (=yhat**) ile **educ** değişkeninin ilişkisini grafik 7.3 de veriyoruz:

Figure 7.3

Estimated relationship between the probability of being in the labor force and years of education, with other explanatory variables fixed.



- Örnekte en çok eğitim alan kadının eğitimi 17 yıldır. EDUC=17 için işgücüne katılma olasılığı 0.5 bulunuyor, *ceteris paribus.*
- Örnekte EDUC değişkeni >5 olduğu için Figure 7.3 ün sağ tarafını kullanacağız. Eğitimin işgücüne katılma olasılığını negatif etkilediği kısım geçerli değildir.
- *nwifeinc*'in katsayısının yorumu : Kocanın yıllık maaşı 10 birim (yani 10,000$)artarsa, karısının işgücüne katılma olasılığı 0.034 düşer.
- Exper karesel biçimde alınmış. İşareti +, karesinin işareti – olduğu için, deneyim işgücüne katılma şansını artırıyor, ancak bu artış giderek azalıyor.

- Beklendiği üzere kadının işgücüne katılma olasılığını belirleyen en önemli faktör küçük çocuğunun olmasıdır. Kidslt6 değişkeninin katsayısı -0.262 dir.
- Yani,*ceteris paribus*, 6 yaşından küçük çocuk sayısında 1 birimlik (1 çocuk) artış kadının işgücüne katılma olasılığını %26.2 azaltmaktadır.
- Örnekte 6 yaşından küçük en az bir çocuğu olan kadınlar toplamın beşte biri kadardır.

LPM modelinin yetersizlikleri

- (7.29) daki LPM modeli dikkat edilebileceği gibi, bağımsız değişkenlerin belli değerleri için sıfırdan küçük ya da birden büyük olasılık tahminleri (inlf hat = yhat) verebilir, ki bu olasılık ilkeleri ile çelişir. Olasılık o ile 1 arasında her zaman negatif-olmayan bir sayıdır.

- (7.29) daki örnekte içerilen 753 kadından sadee 16'sı için inlfhat<0, 17'si için de >1 çıkmıştır.
- Ancak, LPM modelinin asıl yetersizliği bu değildir. Modelde, olasılığın x değerlerine doğrusal bir şekilde bağlı olması büyük mahsur yaratmaktadır.

- Örneğin, (7.29) da bir kadının ilave bir küçük çocuğa sahip olması onun işgücüne katılma şansını %26.2 puan azaltmaktadır.İlişki doğrusal (linear) olduğu için, bu 0.262 rakamı, sıfırıncı çocuktan 1.ci çocuğa geçişte de, diyelim, 3.cü çocuktan 4.cü çocuğa geçişte de aynıdır.
- Oysa bu gerçekçi değildir. İlişki büyük olasılıkla doğrusal-olmayan (nonlinear) niteliktedir. Yani ilk çocukta annenin evde kalma olasılığı 2.ci çocuk olduğunda evde kalma olasılığından daha yüksektir.

- Bu yetersizliklerine rağmen **doğrusal olasılık modeli** (*linear probability model*, **LPM**) faydalı bir modeldir ve iktisatta çokça kullanılmaktadır.
- Özellikle,bağımsız değişkenler (x'ler) kendi ortalamalarına yakın değerler aldıklarında LPM oldukça başarılı tahmin vermektedir.
- Örneğin, (7.29) daki örnekte kadınların %96'sı ya hiç ya da 1 tane küçük çocuğa sahiptir. Dolayısıyla ***kidslt6*** değişkeninin katsayısı (-0.262) aslında bize ilk çocuğun annenin işgücüne katılma şansını ne kadar azalttığını vermektedir. Bu katsayısı, diyelim, 4.cü çocuktan 5.ci çocuğa geçiş vb için kullanmamak lazım.

LPM'in artıklarının varyansı heteroscedastic'dir.

- LPM'de y ikili (binary) değişken olduğu için Gauss-Markov varsayımlarından birisi (*homoscedasticity*) ihlal olmaktadır:
- **Var (u)= Var(y|x) = p(x)*[1-p(x)]** dir.
- Burada "p(x) = βo + β1x1….+ β(k)x(k) + u" olduğu için artıkların varyansı y'nin koşullu beklenen değerine, o da x'lerin alacağı değerlere bağlıdır.
- Heteroscedasticity altında OLS tahmin edicilerin sapmasız olduklarını ancak etkin olmadıklarını (*inefficient*) biliyoruz.

- Bu heteroscedasticity'nin bir sonucu olarak(7.29) daki standart hatalar genellikle geçerli olmayacaktır.Onları kullanırken bu özelliklerini unutmamalı.
- Ancak, çoğu çalışmada LPM katsayılarının standart hatalarının gerçeği oldukça yansıttıkları da gözlemlenmekte ve standart OLS analizi yapılmaya devam edilmektedir.

EXAMPLE 7.12
(A Linear Probability Model of Arrests)

- Arr86 =1 kişi 1986 yılında tutuklanmışsa,

  =0 tutuklanmamışsa.

Örnek (sample) : 1960-61 California doğumlu ve 1986 dan önce en az bir kez tutuklanmış gençler. **CRIME 1.RAW**

pcnv : hükümle sonuçlanan önceki tutuklamaların oranı;***ptime86***: 1986 da yatılan süre;

avgsen: önceki hükümlerden içeride yatılan süre (ay); ***tottime***: 18 yaşından itibaren 1986 ya kadar yatılan hapis süresi. ***qemp86***: 1986 da çalışılan çeyrek-yıl sayısı.

- Tahmin edilen denklem :

$$\widehat{arr86} = \underset{(.017)}{.441} - \underset{(.021)}{.162}\, pcnv + \underset{(.0065)}{.0061}\, avgsen - \underset{(.0050)}{.0023}\, tottime - \underset{(.005)}{.022}\, ptime86 - \underset{(.005)}{.043}\, qemp86 \qquad (7.31)$$

$$n = 2{,}725,\ R^2 = .0474.$$

- Avgsen'in katsayısı + işaret taşıyor. Yani, eski mahkumiyetlerin caydırıcı etkisi yok.
- Tottime anlamsız çıkmış. Bazi diğer çalışmalarda bu katsayı + bulunmuş [Grogger (1991)].

- 1986 da hapiste geçirilen her ay 1986 da tutuklanma olasılığını %2.2 azaltıyor. Tüm yılı içeride geçiren birisinin tutuklanma olasılığı sıfır çıkmalı : Diğer değişkenleri sıfıra eşitlersek, 0.441- 0.022*(12) = 0.177 kalıyor. Yani, “LPM’i x’lerin tüm değerleri için uygulayamayız “ sonucu burada da karşımıza çıkıyor.
- İşte çalışmak (qemp86) tutuklanma olasılığını ciddi bir şekilde düşürüyor : Diğer faktörler sabitken 1986 ‘nın 4 çeyreğinde de çalışmış bir kişinin tutuklanma olasılığı işsiz birisinden -0.043*4 = -0.172 kadar daha düşüktür.

- (7.31) e kukla bağımsız değişkenler ekleyebiliriz. Örneğin ırk kuklaları ekleyelim : *black ve hispan*

$$\widehat{arr86} = .380 - .152\ pcnv + .0046\ avgsen - .0026\ tottime$$
$$(.019)\quad (.021)\qquad (.0064)\qquad (.0049)$$
$$- .024\ ptime86 - .038\ qemp86 + .170\ black + .096\ hispan \qquad \textbf{(7.32)}$$
$$(.005)\qquad (.005)\qquad (.024)\qquad (.021)$$
$$n = 2{,}725,\ R^2 = .0682.$$

- *Baz gruba (beyazlar) göre siyahların tutuklanma olasılığı ceteris paribus %17 puan daha fazladır. İpanyol kökenlilerinki ise %9.6 puan daha fazla.*

LPM'in program değerlendirmede (evaluation) kullanımı

- Herhangi bir programın sonuçlarını değerlendirmek istiyoruz. İki farklı grup (kategori) olacaktır : 1) programa katılanların oluşturduğu grup. Bu gruba deney grubu (***experimental group***) ya da terapi grubu (***treatment group***) diyoruz.
- 2) Programa katılmayan ancak kontrol (mukayese) amacıyla kapsanan grup.Bu gruba kontrol grubu (***control group***) diyoruz.Genellikle grupların seçimi rasgele (random) değildir.

- ***hrsemp***: firma bazında çalışan başına yapılan eğitim (saat); ***grant***, 1988 yılında firmanın herhangi bir eğitim yardımı alıp almadığını gösteren kukla değişken. ***employ*** :çalışan sayısı, ***sales*** : satışlar.
-

EXAMPLE 7.3

(Effects of Training Grants on Hours of Training)

Using the 1988 data for Michigan manufacturing firms in JTRAIN.RAW, we obtain the following estimated equation:

$$\widehat{hrsemp} = \underset{(43.41)}{46.67} + \underset{(5.59)}{26.25}\, grant - \underset{(3.54)}{.98}\, \log(sales) - \underset{(3.88)}{6.07}\, \log(employ) \qquad (7.7)$$

$$n = 105,\ R^2 = .237.$$

- Grant’ın katsayısı anlamlı ve pozitiftir. Eğitim yardımı almış firmalar, ceteris paribus, ortalama olarak 26.25 saat daha fazla eğitim yapmaktadırlar. Hrsemp değişkeninin örnek ortalaması 17, max değeri ise 164 dür. Dolayısıyla, eğitimi büyük ölçüde yardım belirlemektedir.
- Firma satışlarının eğitim çabaları ile ilişki çıkmamaktadır.
- İstihdam düzeyi eğitimle ters yönde ilişkili ve katsayının t değeri -1.56 (=-6.07/3.88), %10 düzeyinde anlamlı.