HAFTA 13

GÖLGE DEĞİŞKENLERLE REGRESYON (DUMMY VARIABLES)

Gölge veya kukla (dummy) değişkenler denen nitel değişkenler, cinsiyet, din, ten rengi gibi hemen sayısallaştırılamayan ama açıklanan değişkenin davranışını etkileyebilen değişkenlerin regresyon modeline alınması *gölge değişkenli regresyon modelini* oluşturur. Böyle $0-1$ değerlerini alan değişkenlere *gölge değişkenler* denir. Gölge değişkenler regresyon modellerinde tıpkı nicel değişkenler gibi kullanılabilir. Regresyon modelindeki değişkenlerin hepsi gölge ya da nitel ise böyle modellere *varyans analizi modelleri* denir.

Örneğin; $Y=$ bir profesörün yıllık maaşı

$$D=\begin{cases}1, & \text{erkek profesör}\\ 0, & \text{bayan profesör}\end{cases}$$

Model: $Y_i=\beta_0+\beta_1 D_i+\varepsilon_i$

Burada cinsiyetin profesör maaşı değişiminde etkisi olup olmadığı araştırılacaktır. Kuşkusuz ki, yaş, akademik derece, kıdem gibi diğer değişkenler sabit tutulacaktır. Hata terimlerinin regresyon varsayımlarını sağladığı koşulu altında

$E(Y_i|D_i=0)=\beta_0$ kadın profesörlerin ortalama maaşı

$E(Y_i|D_i=1)=\beta_0+\beta_1$ erkek profesörlerin ortalama maaşı

cinsiyetin profesör maaşlarına etkisi olup olmadığı $H_0:\ \beta_1=0$ hipotezi ile test edilir.

Örnek: Profesörlerin işe başlama maaşlarına ilişkin veriler

Y (bin dolar)	Cinsiyet (1=erkek, 0=kadın)
22	1
19	0
18	0
21.7	1
18.5	0
21	1
20.5	1
17	0
17.5	0
21.2	1

Kestirim denklemi

$Y_i=\ 18\ +\ 3.28D_i$

$S_\beta:\ 0.32\quad 0.44$

$t:54.74\quad 7.439$

R^2:0.8737

$\hat{\beta}_0=1800$ kadın profesörlerin tahmin edilen ortalama maaşı

$\hat{\beta}_0+\hat{\beta}_1=21280$ erkek profesörlerin tahmin edilen ortalama maaşı

Sonuç olarak, kadın profesörlerin ortalama maaşı erkek profesörlerinkinden düşüktür.

Nicel ve nitel değişkenlerin karışık olduğu regresyon modellerine *ortak varyans çözümlemesi modelleri* denir.

Biri nicel, biri iki değerli nitel değişkenli regresyon:

Model: $Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \varepsilon_i; \quad \varepsilon_i \sim N\left(0, \sigma^2\right)$

Bir önceki örneğe dönersek,

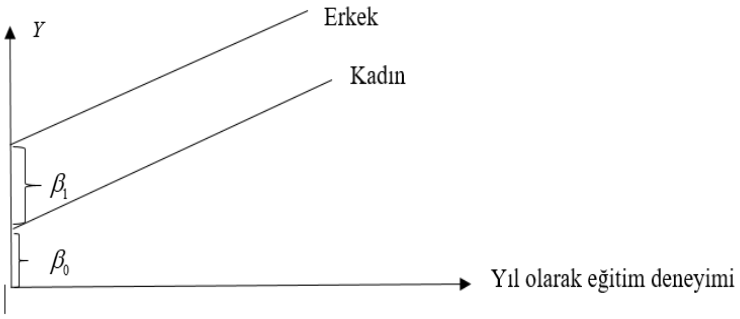
$Y_i =$ bir profesörün yıllık maaşı

$X_i =$ yıl olarak eğitim deneyimi

$$D_i = \begin{cases} 1, & \text{erkek} \\ 0, & \text{kadın} \end{cases}$$

Bir kadın profesörün ortalama maaşı: $E\left(Y_i \middle| X_i, D_i = 0\right) = \beta_0 + \beta_2 X_i$

Bir erkek profesörün ortalama maaşı: $E\left(Y_i \middle| X_i, D_i = 1\right) = \left(\beta_0 + \beta_1\right) + \beta_2 X_i$



Böyle bir regresyon modelinde dikkat edilmesi gereken özellikler:

1. Kadın ve erkek gibi iki grubu belirlemek için bir gölge değişken yerine iki gölge değişken tanımlanırsa model

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 X_i + \varepsilon_i; \quad \varepsilon_i \sim N\left(0, \sigma^2\right)$$

$$D_{1i} = \begin{cases} 1, & \text{erkek} \\ 0, & \text{değil} \end{cases} \qquad D_{2i} = \begin{cases} 1, & \text{kadın} \\ 0, & \text{değil} \end{cases}$$

bir önceki örnekten Y_i ve X_i tanımlanırsa, bu model için tasarım matrisi

$$X = \begin{matrix} \beta_0 & \beta_1 & \beta_2 & \beta_3 \end{matrix} \begin{bmatrix} 1 & 1 & 0 & x_1 \\ 1 & 1 & 0 & x_2 \\ 1 & 0 & 1 & x_3 \\ 1 & 1 & 0 & x_4 \\ 1 & 0 & 1 & x_5 \end{bmatrix}$$

dir. Görüleceği üzere 2. ve 3. sütunun toplamı 1. sütunu vermektedir. D_{1i} ve D_{2i} arasında tam bir ortak doğrusallık vardır ve EKK tahmini yapmak olanaksızdır. Böyle durumlarda iki gölge değişken yerine bir gölge değişken kullanmak ortak doğrusallık sorununu

çözecektir. O halde bir nitel değişkende m öbek varsa, yalnızca $m-1$ gölge değişken kullanmakla gölge değişken tuzağı olarak belirtilen ortak doğrusallık sorunundan kurtulmayı sağlar.

2. Gölge değişken kullanan regresyon modelleri yorumlanırken $0-1$ değişkenlerinin nasıl verildiği önemlidir.

3. 0 değeri verilen öbek, şık yada düzeye temel şık, ölçü şıkkı, kontrol şıkkı, karşılaştırma şıkkı, başvuru şıkkı yada atlanan şık gibi adlar verilir. Bu şık öbürlerinin karşılaştırılmaları için bir temeldir. Hangi şıkkın temel şık olacağı önsel bazı düzencelerin etkili olduğu bir seçimden başka bir şey değildir.

4. D gölge değişkenine verilen β_1 katsayısı sabit terim farkı olarak adlandırılır. 1. değerini alan şıkkın sabit teriminin temel şıkkın sabit terim katsayısından ne kadar farklı olduğunu gösterir.

Biri nicel, biri ikiden çok değer alan nitel değişkenli regresyon:

Bir kimsenin yıllık sağlık harcamalarının, o kimsenin gelirine ve eğitimine göre regresyon modeli bulunmak istenirse, değişkenler

$Y_i =$ yıllık sağlık harcaması
$X_i =$ yıllık gelir
$D_i = \{\text{orta ögretim, lise, üniversite}\}$

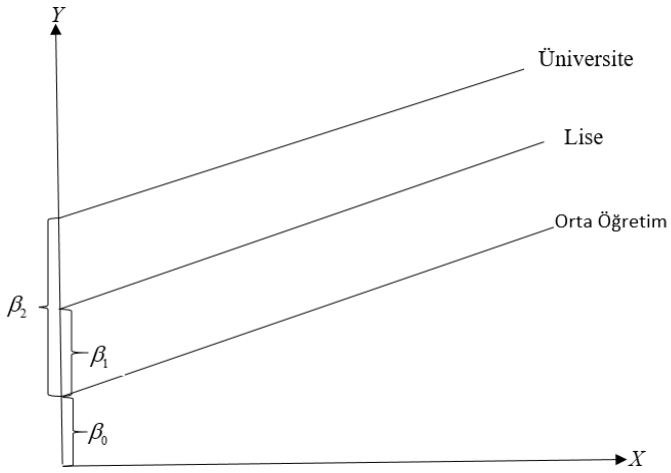
dir. Gölge değişken sayısı, değişken düzey sayısından bir eksik olmalı kuralı gereğince D_1, D_2, D_3 gölge değişkenleri yerine modele D_1 ve D_2 gölge değişkenleri alınır.

$$D_1 = \begin{cases} 1, & \text{lise mezunu} \\ 0, & \text{değil} \end{cases} \qquad D_2 = \begin{cases} 1, & \text{üniversite mezunu} \\ 0, & \text{değil} \end{cases}$$

gölge değişkenleri alınarak, model

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 X_i + \varepsilon_i; \quad \varepsilon_i \sim N\left(0, \sigma^2\right)$$

olup, ortaöğretim düzeyi keyfi olarak temel düzey alınır. Öyleyse β_0 sabit terimi bu düzeyin sabitidir.



$E\left(Y_i \middle| X_i, D_{1i}=0, D_{2i}=0\right)=\beta_0+\beta_3 X_i$ ortaöğretim

$E\left(Y_i \middle| X_i, D_{1i}=1, D_{2i}=0\right)=\left(\beta_0+\beta_1\right)+\beta_3 X_i$ lise

$E\left(Y_i \middle| X_i, D_{1i}=0, D_{2i}=1\right)=\left(\beta_0+\beta_2\right)+\beta_3 X_i$ üniversite

Regresyon modeli bulunduktan sonra β_1 ile β_2 fark sabit terimlerinin tekil olarak, temel düzeyden, istatistik bakımından anlamlı bir fark gösterip göstermediği $H_0: \beta_1=\beta_2=0$ hipoteziyle test edilebilir. ANOVA ve ANCOVA ile de test edilebilir.

Farklı bir gölge değişken tanımlama yolu kullanıldığında regresyon modelinin yorumlanması da değişecektir.

Biri nicel, ikisi nitel değişkenli regresyon:

Gölge değişken tekniği birden çok nitel değişken için genişletilebilir. Profesör maaşları örneğine dönecek olursak,

$Y_i=$ bir profesörün yıllık maaşı
$X_i=$ yıl olarak eğitim deneyimi

$$D_{1i}=\begin{cases}1, & \text{erkekse}\\ 0, & \text{değilse}\end{cases} \qquad D_{2i}=\begin{cases}1, & \text{beyazsa}\\ 0, & \text{değilse}\end{cases}$$

Artık atlanan ya da temel şıkkı ten rengi yani burada zenci kadın profesörse

Model: $Y_i=\beta_0+\beta_1 D_{1i}+\beta_2 D_{2i}+\beta_3 X_i+\varepsilon_i; \quad \mathrm{E}\left(\varepsilon_i\right)=0$

Zenci kadın profesörün ortalama maaşı: $E\left(Y_i \middle| D_{1i}=0, D_{2i}=0, X_i\right)=\beta_0+\beta_3 X_i$

Zenci erkek profesörün ortalama maaşı: $E\left(Y_i \middle| D_{1i}=1, D_{2i}=0, X_i\right)=\left(\beta_0+\beta_1\right)+\beta_3 X_i$

Beyaz kadın profesörün ortalama maaşı: $E\left(Y_i \middle| D_{1i}=0, D_{2i}=1, X_i\right)=\left(\beta_0+\beta_2\right)+\beta_3 X_i$

Beyaz erkek profesörün ortalama maaşı: $E\left(Y_i \middle| D_{1i}=1, D_{2i}=1, X_i\right)=\left(\beta_0+\beta_1+\beta_2\right)+\beta_3 X_i$

Yukarıdaki modellerde sabit terimleri farklı almakla birlikte β_3 eğim katsayıları aynıdır.

Regresyon parametrelerinin EKK tahmin edicilerinden β_2 istatistiksel anlamlı ise ten rengi, β_1 istatistiksel anlamlı ise cinsiyet bir profesörün maaşını etkiliyor demektir. Eğer β_1 ve β_2'nin her ikisi de istatiksel anlamlı ise hem cinsiyet hem de ten rengi profesörlerin maaşlarında önemli birer belirleyicidir.

Regresyon modellerinin kararlılıklarının sınaması:

Şimdiye kadar alınan modellerde nitel değişkenlerin çeşitli alt regresyonlarında sabit terimi etkilediği ama eğim katsayıları aynı kaldığı varsayıldı. Nitel değişkenin her düzeyi için farklı bir regresyon doğrusu elde edilirken bu doğruların aynı eğimli paralel olduğu incelendi. Eğer bu doğruların eğimleri farklıysa, sabit terimlerinin sınanmasının uygulamada anlamı kalmaz. Bu regresyon doğrularının farklı eğimli olup olmadığına çeşitli testlerle bakılabilir.

Örnek: $1946-1963$ yıllarında İngiltere'de tasarruflar ve gelir verileri iki dönemde incelenecektir.

I. dönem: Yeniden yapılanma 1946 – 1954 arası (II. Dünya savaşı sonrası)
II. dönem: Yeniden yapılanma sonrası 1955 – 1963 arası

I. dönem: $Y_i = \lambda_1 + \lambda_2 X_i + u_{1i};\ \ i = 1, 2, \ldots, n_1$
II. dönem: $Y_i = \gamma_1 + \gamma_2 X_i + u_{2i};\ \ i = 1, 2, \ldots, n_2$

$Y =$ tasarruflar (milyon \$)

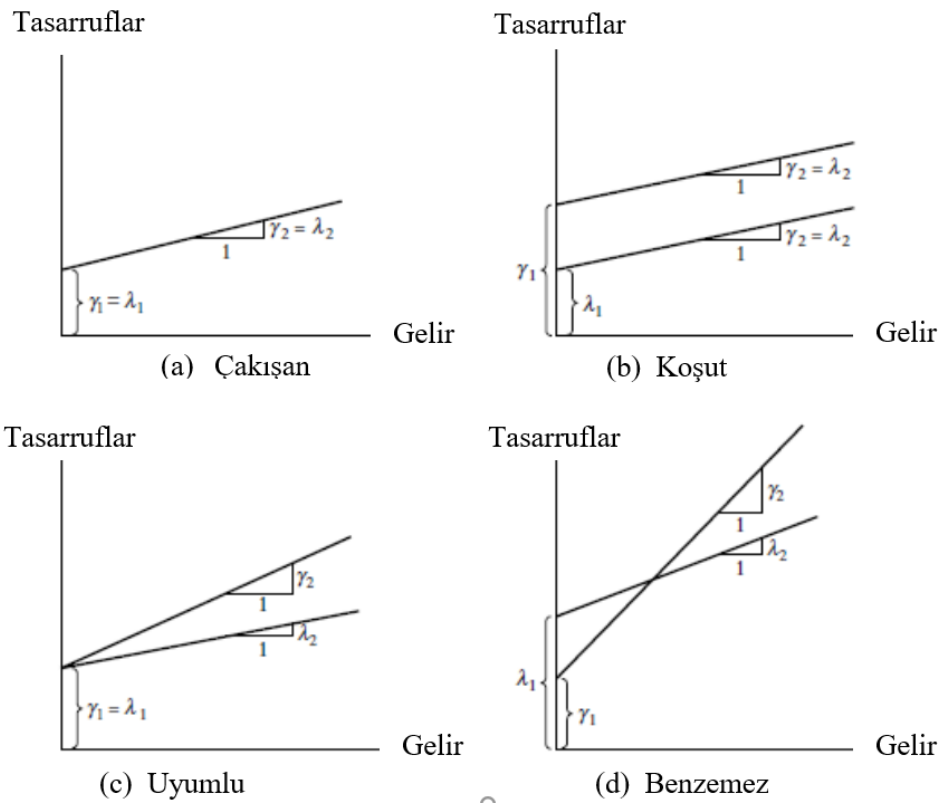
$X =$ gelir (milyon \$)

$u_{1i},\ u_{2i} =$ iki regresyon modelindeki hata terimleri

n_1 ve n_2 sırasıyla I. ve II. dönemdeki gözlem sayıları

Her iki dönem için regresyon doğruları:

1. $\lambda_1 = \gamma_1$ ve $\lambda_2 = \gamma_2$ ise çakışan regresyonlar (aynı)
2. $\lambda_1 \neq \gamma_1$ ve $\lambda_2 = \gamma_2$ ise koşut regresyonlar (konumları bakımından farklı)
3. $\lambda_1 = \gamma_1$ ve $\lambda_2 \neq \gamma_2$ ise uyumlu regresyonlar (aynı sabit terimli fakat farklı eğimli)
4. $\lambda_1 \neq \gamma_1$ ve $\lambda_2 \neq \gamma_2$ ise benzemez regresyonlar



Her iki dönem için regresyonlar ayrı ayrı bulunabilir, sonra yukarıdaki durumların her biri sınanabilir.

Chow sınaması:

Veride yapısal bir değişimin olup olmadığı Gregory Chow'un önerdiği Chow sınaması ile test edilebilir.

Chow sınamasının varsayımları:

a) $u_{1i} \sim N\left(0, \sigma^2\right)$ ve $u_{2i} \sim N\left(0, \sigma^2\right)$ aynı varyanslı

b) u_{1i} ve u_{2i} bağımsız rasgele değişkenler

Chow sınamasının adımları:

1. n_1 ve n_2 gözlemleri birleştirilerek tek bir regresyon doğrusundan hata terimleri tahmin edilir (artıklar) ve bu artıklarda elde edilen $SSE = S_1$ bulunur.

$n = n_1 + n_2 =$ toplam gözlem sayısı,

$p = (k+1) =$ modeldeki parametre sayısı

$k =$ açıklayıcı değişken sayısı

$n - p = n_1 + n_2 - p =$ serbestlik derecesi

2. Daha sonra n_1 gözlem ve n_2 gözlem için ayrı ayrı regresyon doğrularından hata terimleri tahmin edilir. Her bir kestirim denkleminden $SSE_1 = S_2$ ve $SSE_2 = S_3$ elde edilir.

I. dönem: $sd_1 = n_1 - p$

II. dönem: $sd_2 = n_2 - p$

Her iki SSE_1 ve SSE_2 toplanır ve $S_2 + S_3 = S_4$ bulunur. Burada serbestlik derecesi $sd_3 = n_1 + n_2 - 2p$ dir.

3. $S_5 = S_1 - S_4$ bulunur ve serbestlik derecesi ise

$sd_4 = (n_1 + n_2 - p) - (n_1 + n_2 - 2p) = -p + 2p = p$ dir.

4. Chow sınaması varsayımları altında önerilen *F* test istatistiği

$$F = \frac{S_5 / sd_4}{S_4 / sd_3} = \frac{(S_1 - S_4)/p}{(S_2 + S_3)/(n_1 + n_2 - 2p)} \sim F_{p;n_1+n_2-2p}$$

dir ve yokluk hipotezi

H_0 : Her iki regresyon aynıdır. $(\beta_0 = \gamma_0$ ve $\beta_1 = \gamma_1)$

hipotezini test eder.

Örnek: Bir önceki örnekten, $n_1 = 9,\ n_2 = 9,\ p = 2$

1. Tüm veri için:

$\hat{Y}_i = -1.0821 + 0.1178 X_i$

$S_{\hat{\beta}}$: 0.1452 0.0088

t : –7.4548 13.4316

$r^2 = 0.9185$ ve $S_1 = SSE = 0.5722 \quad sd = 16$

2. I. dönem

$\hat{Y}_i = -0.2622 + 0.0470 X_i$

$S_{\hat{\lambda}}$: 0.3054 0.0266

t : –0.8719 1.7700

$r^2 = 0.3095$ ve $S_2 = SSE_1 = 0.1396 \quad sd_1 = 7$

II. dönem

$\hat{Y}_i = -1.7502 + 0.1504 X_i$

S_{γ} : 0.3567 0.0175

t : –4.8948 8.5749

$r^2 = 0.9131$ ve $S_3 = SSE_2 = 0.1931 \quad sd_2 = 7$

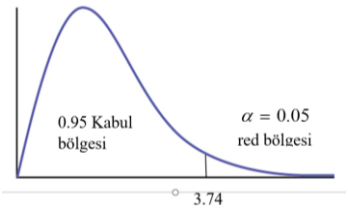
3. $S_4 = S_2 + S_3 = 0.1396 + 0.1931 = 0.3327, \quad sd_3 = 14$

$S_5 = S_1 + S_4 = 0.5722 - 0.3327 = 0.2395, \quad sd_4 = 2$

4. Test istatistiği

$$F = \frac{S_5 / sd_4}{S_4 / sd_3} = \frac{0.2395/2}{0.3327/14} = 5.04$$

$\alpha = 0.05$ için $F^*_{2,14}(0.05) = 3.74$



5.04>3.74 olduğundan $H_0: \lambda_1 = \gamma_1$ ve $\lambda_2 = \gamma_2$ hipotezi red edilir.

Yorum: Tasarruf fonksiyonu her iki dönem için farklıdır. Acaba bu fark sabit terimlerden mi yoksa eğimlerden mi olduğunu saptamak için Chow sınaması uyarlanabilir. Aynı zamanda gölge değişkenler yoluyla da bu araştırılabilir.

Gölge değişken yaklaşımı ile iki regresyon karşılaştırılması:

Chow sınaması süreci gölge değişken tekniği ile önemli ölçüde kısıtlanabilir. Uygulamada Chow ve gölge değişken sınamalarından aynı sonuçlar elde edildiyse, gölge değişkenlerin bazı üstünlükleri vardır.

Tasarruf – gelir örneğine dönersek;

$$D_i = \begin{cases} 1, & i. \text{ veri I. dönemde ise} \\ 0, & i. \text{ veri II. dönemde ise} \end{cases}$$

her iki dönem birleştirilerek, regresyon modeli

$$Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 X_i + \alpha_3 D_i X_i + \varepsilon_i$$

I. dönem: $E\left(Y_i \middle| X_i, D_i = 0\right) = \alpha_0 + \alpha_2 X_i$

II. dönem: $E\left(Y_i \middle| X_i, D_i = 1\right) = \left(\alpha_0 + \alpha_1\right) + \left(\alpha_2 + \alpha_3\right) X_i$

Önceki tanımlamadan

$\lambda_1 = \alpha_0,\ \lambda_2 = \alpha_2$ ve $\gamma_1 = \alpha_0 + \alpha_1,\ \gamma_2 = \alpha_2 + \alpha_3$

dir. Burada $\alpha_1 =$ sabit terim farkı ve $\alpha_3 =$ eğim farkı katsayılarını göstermektedir.

Gölge değişken D'nin çarpım $(D * X)$ kalıbında modele eklenmesi iki dönem eğim katsayılarının farklı olup olmadığını ortaya çıkarır.

Tasarruf – gelir modelinin kestirim denklemi:

$$\hat{Y}_i = -1.7502 + 1.4839 D_i + 0.1504 X_i - 0.1034 D_i X_i$$

$S_{\hat{\alpha}}$: 0.3319 0.4704 0.0163 0.0332

t: -5.2733 3.1545 9.2238 -3.1144

$\bar{R}^2 = 0.9425$

Görüleceği üzere hem sabit terim farkı hem de eğim farkı katsayısı istatistik bakımından anlamlıdır. Bu da her iki dönem için öngörülen regresyon modellerinin farklı olduğunun güçlü bir göstergesidir.

$D_i = 1$ ise I. dönem: $\hat{Y}_i = (-1.7502 + 1.4839) + (0.1504 - 0.1034) X_i$

$= -0.2663 + 0.0470 X_i$

$D_i = 0$ ise II. dönem: $\hat{Y}_i = -1.7502 + 0.1504 X_i$

kestirim denklemleri Chow sınaması ile bulunan kestirim denklemleri aynıdır.

Gölge değişken tekniğinin üstünlükleri:

1. Yalnızca tek bir regresyon modeli bulmak yeterlidir. Tekil regresyonlar buradan türetilebilir.
2. Tek regresyon modelindeki regresyon parametrelerinin testlerinin yapılması ile tekil regresyonların farklı olup olmadıkları bulunabilir.
3. Chow sınaması tekil regresyonların sabit terimleri mi yoksa eğimleri açısından farklı olup olmadıklarının ayrımını yapamaz. Buna karşın gölge değişken tekniği Chow sınamasına karşın üstünlük sağlar.
4. Verilerin bir araya getirilmesi serbestlik derecesini yükseltip tahmin edilen ana kütle katsayılarının göreli hassaslığını artırır.

İki gölge değişkenin etkileşimi

$Y_i =$ bir profesörün yıllık maaşı

$X_i =$ yıllık gelir

$$D_{1i} = \begin{cases} 1, & \text{kadınsa} \\ 0, & \text{erkekse} \end{cases} \qquad D_{2i} = \begin{cases} 1, & \text{üniversite mezunuysa} \\ 0, & \text{değilse} \end{cases}$$

Kadın üniversite mezunu, erkek üniversite mezununa göre giyim için daha fazla harcama yapılabilir. Yani; iki gölge değişken arasında etkileşim olabilir. Bunu anlamak için model;

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 \left(D_{1i} D_{2i} \right) + \beta_4 X_i + \varepsilon_i$$

$$E\left(Y_i \middle| D_{1i} = 1,\ D_{2i} = 1,\ X_i \right) = \left(\beta_0 + \beta_1 + \beta_2 + \beta_3 \right) + \beta_4 X_i$$

$\beta_1 =$ kadın olmanın fark etkisi

$\beta_2 =$ üniversite mezunu olmanın fark etkisi

$\beta_3 =$ kadın üniversite mezunu olmanın fark etkisi

Kadın üniversite mezunlarının ortalama giyim harcamasının, kadınların ya da üniversite mezunlarının ortalama giyim harcamasından β_3 kadar farklı olduğunu gösterir. Eğer $\beta_1,\ \beta_2,\ \beta_3$'ün hepsi artı işaretliyse, kadınların ortalama giyim harcaması daha yüksektir, ama

bu kadınlar bir de üniversite mezunuysa bu fark daha da büyüktür. Benzer bir biçimde bir üniversite mezununun ortalama giyim harcaması, temel öbekten daha yüksektir, ama bu üniversite mezunu bir de kadınsa bu fark daha da artar. Buda bize, etkileşim gölge değişkeninin iki farklı özelliğinin tekil etkilerini nasıl değiştirdiğini gösterir. Etkileşim gölge değişken katsayısının istatistik bakımından anlamlı olup olmadığı alışıldık t sınamasıyla sınanabilir. Eğer anlamlı çıkarsa iki özelliğin birlikte varlığı bu özelliklerin tek tek etkilerini azaltacak ya da artıracak demektir. Önemli bir etkileşim terimini göz ardı etmenin model kurma hatasına yol açacağını söylemeye bile gerek yoktur.