Öneri. C.1. S.6
Ocak 1997. ss.99-107.

GÖLGE DEĞİŞKENLER İLE REGRESYON

Dr.Talat FIRLAR

İstanbul Üniversitesi, Meslek Yüksek Okulu, Yardımcı Doçent

ABSTRACT: In this paper; the general properties and some of the benefits of structured approach for dummy variables development are summarized and structured tools and methodologies are reviewed.

Bu çalışmada, regresyon analizindeki kalitatif degişkenlerin rolünün açıklanmasına çalışılmaktadır. Kalitatif değişkenler gölge değişken (dummy variables) olarak adlandırılır. Gözlemsel çalışmalardaki birçok ilginç problemin çözümü için dogrusal regresyonda oldukça esnek bir araç olarak gölge değişkenleri kullanırız.

I-GÖLGE DEGİŞKENLERİN YAPISI

Regresyon analizindeki değişkenler. yalnızca kantitatif (sayısal olarak ifade edilebilen degişkenler. girdi. çıktı. fiyat. maliyet. yükseklik. sıcaklık v.b.) değil kalitatif değişkenler de (cinsiyet. soy. renk. din. milliyet. savaşlar. depremler. grev. politik kargaşa ve ekonomik politikalardaki değişmeler gibi) olabilir. Örneğin. bütün diğer faktörleri sabit kabul ederek. kadın kolej öğretmenlerinin diğer cinsiyettekilere göre daha az kazandığını bularak ve beyaz olmayanların beyaz olanlardan daha az kazandığını bularak cinsiyet ve ırk ayrımı yapıldığı şeklinde bir sonuca varabiliriz. Sonuç ne olursa olsun cinsiyet ve ırk gibi kalitatif değişkenler bağımlı değişkeni etkiler ve bu değişkenler (kalitatif değişkenler) açıklayıcı değişkenler olarak kabul edilir[1].

Kalitatif değişkenler genellikle bir kalite veya bir özelliğin olup olmadığını gösterir. Erkek veya kadın. siyah veya beyaz. katolikler veya katolik olamyanlar gibi. Böylesi özellikleri bir yöntemle göstermek istersek 1 ve 0 gibi yapay değişkenler kullanarak o özelliğin olup olmadığını belirtiriz. 1 o özelliğin olduğunu. 0 ise o özelliğin olmadığını gösterir. Örneğin 1 erkekleri. 0 kadınları . 1 kolej mezunu olanları. 0 kolej mezunu olmayanları gösterebilir. 1 ve 0 atanan değişkenlere gölge değişken denir. Gölge değişkenler için gösterge değişkenler. ikili değişkenler. kategorik değişkenler ve dikotomus değişkenleribi adlar da kullanılabilir.

Bir regresyon modeli açıklayıcı değişken olarak yalnızca gölge değişkenleri içerebilir. Böyle modellere Varyans Analizi Modelleri (AOV) denir. Böylesi bir modele örnek olarak :

$Y_i = \alpha + \beta . D_i + U_i$
Y = Bir kolej öğretmeninin yıllık kazancı
$D_i = 1$ Erkek kolej öğretmeni
$U_i = 0$ Diğerleri (Kadın kolej öğretmeni)

Bu model iki değişkenli regresyon modeline benzemektedir. X değişkeni yerine D gölge değişkeni kullanılmaktadır. (Gölge değişkeni D ile göstereceğiz.) Bu model bize (yaş . derece ve çalışma yılı gibi diğer değişkenler sabit kalmak koşuluyla) bir kolej ögretmeninin maaşındaki farkı cinsiyeti ile açıklayabilir. Klasik bir doğrusal regresyon modeli olarak düşünürsek :

Kadın kolej öğretmenlerinin ortalama maaşı
$E(Y_i | D_i=0) = \alpha$
Erkek kolej öğretmenlerinin ortalama maaşı
$E(Y_i | D_i=1) = \alpha + \beta$

Başlangıç terimi α . kadın kolej öğretmenlerinin ortalama maaşını ve eğim katsayısı β . erkek kolej öğretmenlerinin ortalama maaşının kadın kolej öğretmenlerinkinden farkını gösterir.

$\alpha + \beta$ ise kadın kolej öğretmenlerinin ortalama maaşını gösterir. Bu durumda bir test kurmak gerekirse : $H_0 : \beta = 0$ cinsiyetler arası bir fark olmadığını gösteren hipotez olur. t testi kullanılarak β'nın anlamlı olup olmadığı belirlenir.

Aşağıdaki tabloda cinsiyete göre 10 kolej öğretmeninin başlangıç maaşları verilmektedir.

Ücret (Bin $)	22.0	19.0	18.0	21.7	18.5	21.0	20.5	17.0	17.5	21.2
Cinsiyet (1=erkek 0 kadın)	1	0	0	1	0	1	1	0	0	1

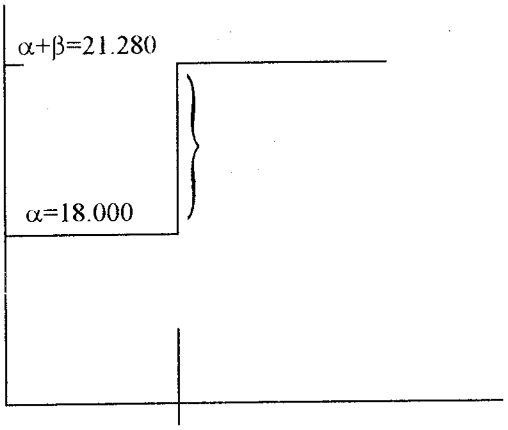
Sonuçta elde edilen regresyon modeli:

$$
\begin{aligned}
Y_i &= 18.00 + 3.28\,D_i \\
&\quad\ (.32) \quad\ (.44) \\
t &= (57.74) \quad (7.439) \\
R^2 &= .8737
\end{aligned}
$$

Sonuc olarak. kadın kolej öğretmenlerinin tahmin edilen ortalama maaşı 18.000$ ve erkek öğretmenlerin tahmin edilen ortalama maaşı 21.280 $ dır. Bu değerler kolaylıkla tablodan erkek ve kadın öğretmenlerin maaşlarının ortalaması olarak bulunabilir.

β istatistik olarak anlamlıdır. Yani iki sınıfın (kategori) maaş ortalamaları farklıdır. Güncel olarak kadın öğretmenlerin ortalama maaşı erkeklerin ortalama maaşından düşüktür.

Aşağıdaki şekilde regresyonun grafiği verilmiştir. bu şekilde veriler erkek ve kadın öğretmenler olarak iki gruba ayrılmıştır. Şekilde görüldüğü gibi regresyon fonksiyonu adım (step) fonksiyonudur. Kadın öğretmenlerin maaş ortalaması 18.000 \$ ve erkek öğretmenlerin maaş ortalaması β = 3.28 \$ olarak 21.280 \$ 'a yükselir.



Kadın Öğretmenler Erkek Öğretmenler

Şekil :1

AOV (Varyans Analizi Modelleri) hernekadar sosyaloji, psikoloji, eğitim ve pazar araştırmalarında kullanılırsa da ekonomik çalışmalarda yaygın olarak kullanılmaz. Tipik olarak ekonomik araştırmlarda kullanılan regresyon modelleri kalitatif ve kantitatif olmak üzere açıklayıcı değişkenler içerir. Kalitatif ve kantitatif değişkenleri karışık olarak içeren regresyon modelleri Kovaryans Analizi Modelleri olarak adlandırılır. Bu metinde bu tip modeller ile çalışacagız.

II-İKİ SINIF VEYA KATEGORİ İLE BİR KANTİTATİF VE BİR KALİTATİF DEĞİŞKEN ÜZERİNDE REGRESYON

ACOV modeline bir örnek olarak aşagıdaki modeli verebiliriz:

$Y_i = \alpha_1 + \alpha_2 D_i + \beta X_i + U_i$

Yi = Bir kolej öğretmeninin yıllık kazancı

Xi = Yıl sayısı olarak deneyim

Di = 1 Erkek Di = 0 Diğer

Bu model bir kantitatif değişken (yılsayısı olarak deneyim) ve bir kalitatif değişken (cinsiyet) içerir.

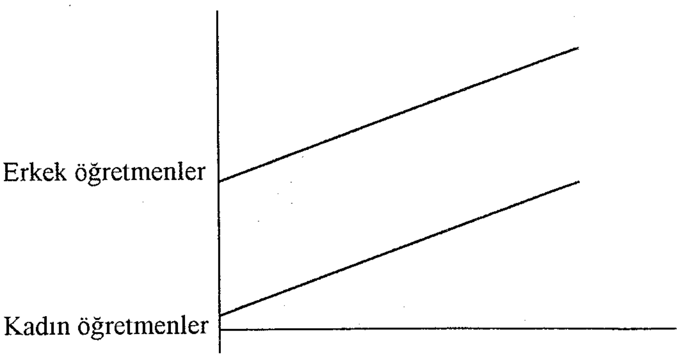
$E(U_i) = 0$ varsayımı ile:

Kadın öğretmenlerin ortalama kazancı :

$E(Y_i \mid X_i, D_i = 0) = \alpha_1 + \beta Xi$

Erkek öğretmenlerin ortalama kazancı :

$E(Y_i \mid X_i, D_i = 1) = (\alpha_1 + \alpha_2) + \beta Xi$



Deneyim Yılı

Şekil : 2

Modelin postülası olarak, erkek ve kadın öğretmenlerin yıllık kazancı ile deneyimleri arasındaki ilişkinin eğimi (β) aynı fakat başlangıç noktaları değişik diyebiliriz. Bir başka deyişle erkek öğretmenlerin kazanç ortalaması kadın öğretmenlerin kazanç ortalamasına göre (α_2) kadar farklı, fakat değişim oranları aynı diyebiliriz.

Genel eğimin tahmini geçerliyse, iki regresyon için hipotez testi aynı başlangıç noktasını verecek (cinsiyetler arası bir ayrım yok) ve bu, başta verilen regresyon modeli ile kolayca çalıştırılıp tahmin edilen α_2 'nin anlamlı olamdığı t testi ile görülecektir. Eğer t testi α_2 ' nin anlamlı olduğunu gösteriyorsa, erkek ve kadın öğretmenlerin yıllık maaşlarının ortalamasının aynı olduğunu söyleyen boş hipotez testi rededilecektir.

Gölge değişken regresyonunun özellikleri aşağıda sıralanmaktadır:

1. Erkek ve kadın olarak iki kategoriye ayırıyoruz. D_i olarak tek gölge değişkenimiz var. Eğer D_i =1 ise erkek , D_i = 0 ise kadın olduğunu biliyoruz. Yani iki çıktımız var. Böylece bir gölge değişken 2 kategori için yeterli oluyor. Regresyon modelinin bir başlangıç terimi içerdiğini varsayıp modeli yazarsak :

$Y_i = \alpha_1 + \alpha_2 + \alpha_3 D_{3i} + \beta X_i + U_i$

Y_i ve X_i daha önce tanımlanmıştı.

D_{2i} = 1 Erkek öğretmen

= 0 Diğer

D_{3i} = 1 Kadın öğretmen

= 0 Diğer

Modelde D_2 ile D_3 arasında çok fazla bir doğrusal bağımlılık (collinearity) olğundan tahmin yapılamaz.

Bunu görmek için 3 erkek öğretmen ve 2 kadın öğretmenden oluşan bir örnek alalım:

	D_2	D_3	X
Erkek $Y_1=1$	1	0	X_1
Erkek $Y_2=1$	1	0	X_2
Erkek $Y_3=1$	0	1	X_3
Erkek $Y_4=1$	1	0	X_4
Erkek $Y_5=1$	0	1	X_5

Tablo :1

İlk kolon α_1 'I (genel başlangıç noktası terimi) gösteriyor. Kolayca görüleceği gibi $D_2 = 1- D_3$ veya

$D_3 = 1-D_2$, yani D_2 ve D_3 tamamen doğrusal bağımlı. Tamamen multicollinearity durumlarında OLS tahminini kullanmak olanaksızdır. Bu sorunu çözmek için değişik yollar vardır. En kolayı, eğer kalitatif değişkenin iki sınıfı veya düzeyi varsa , önceki modelde yaptığımız gibi yalnızca bir gölge değişken kullanmak olacaktır. Bu durumda D_3 kolonuna gerek kalmayacak ve multicollinearity sorunundan kaçılacaktır. Bu durumu genel olarak ifade edersek: Eğer bir kalitatif değişken m kategoriye sahipse yalnızca m-1 gölge değişken girilir. Örneğimizde cinsiyet 2 kategoriye ayrılıyor ve bir gölge değişkenimiz var. Bu kuralı izlemesssek "gölge değişken tuzağına " düşeriz , bu da tam multicollinearrity'dir.

2. 1 ve 0 , erkek ve kadın gibi iki kategoriye atanır. Örneğimizde D=1 kadın ve D=0 erkek olarak atayabiliriz. Bu durumda iki regresyon modeli elde edilir.

Kadın öğretmen $E(Y_i \mid X_i , D_i =1) = (\alpha_1 + \alpha_2) + \beta X_i$

Erkek öğretmen $E(Y_i \mid X_i , D_i =0) = \alpha_1 + \beta X_i$

Modeldeki α_2 , bir kadın öğretmenin maaşının ortalaması bir erkek öğretmenin maaşının ortalamasından ne kadar farklı olduğunu söyler. Bu durumda cinsiyetler arasında fark varsa α_2 'nin (daha önce pozitif olamsı beklenmesine karşın) negatif olması beklenir. Bu yüzden modelin sonuçlarını yorumlarken kullanılan gölge değişkenlere nasıl 1 ve 0 atandığı önemlidir.

3. Grup, kategori veya sınıflamada 0 değeri atananlar çogunlukla temel, kontrol, karşılaştırma veya elde değeri olarak anılır. Bu durumda kategorileri karşılaştırmak için bir temel oluşturur. Böylece ilk örnek verilen modelde kadın öğretmenler temel kategoridir. Başlangıç terim α_1 . regresyonu D=0 ile çalıştırdığımızda temel kategori için başlangıç terimidir. Hangi kategorinin temel kategori olacagı bazen öncelikle belirlenmesi gereken bir seçim sorunudur.

4. Gölge değişken D 'ye bağlanan α_2 katsayısı , "fark başlangıç katsayısı "olarak adlandırılır. Çünkü bu kaytsayı bize 1 ile belirlenen kategorinin başlangıç terminin değerinin, temel kategorinin başlangıç katsayısından farkının ne kadar olacagını gösterir.

Dan M. Bechter ve Stephen H. Pollock aşağıdaki tahmin modelini, 1967-IV 'den 1979-IV 'e kadar Amerikan ekonomisinin tüm ticari sektöründeki envanter dalgalanmasını açıklamak için kurmuşlardır. (t değerleri parantez içindedir.)

$$(t/s) = 1.269 - 0.3615.C + 0.0215.S^e - 0.0227.S - 0.2552.U + 0.0734.DUM$$

$$(19.6) \quad (-2.2) \quad (5.7) \quad (-2.4) \quad (-2.4) \quad (4.8)$$

$$R^2 = 0.71 \quad d=1.91$$

(t/s) : Sabit dolar değeri ile envanterler.

C : 4'den 6 aya kadar olan oranların bir yıl önceki tüketici fiyat indeksindeki yüzde değişiminden küçük olanlar.

S^e : Bulunulan dönemdeki beklenen satışlar. Bu satışlar geçmiş dönem satışlarının sapmalarına bakılarak trendle bulunulabilir.

U : Trend etrafındaki satışların uçuculuğu tarafından ölçülen satışlardaki belirsizlik.

DUM: Gölge değişken.1967-IV 1974-I 0 ve 1974-II 1979-IV 1 olarak.

Hernekadar katsayı anlamlı ve beklenen işareti alıyorsa da gölge değişken üzerinde durmak gerekir. Sonuç bize gösteriyorki; envanter satış oranı 1974 gerileme periyodundan sonra (önceki periyoda göre) daha yüksektir. (=1.2690+0.0734) Böylece regresyon çizgisi bir önceki periyoda göre daha yüksek bir düzeydedir.

III-İKİDEN FAZLA SINIF İLE BİR KANTİTATİF VE BİR KALİTATİF DEĞİŞKEN ÜZERİNDE REGRESYON

Yıllık sağlık harcamalarının , kişinin geliri ve eğitim düzeyine bağlı olduğunu varsayalım. Eğitim kalitatif bir değişken olarak 3 düzeyde gösterilebilir: lise altı, lise ve kolej. Önceki durumdan farklı olarak burada değişkenimiz 2 kategoriden fazla kategoriye sahip. Kural olarak hatırlayacağımız gibi gölge değişken sayısı kategori sayısının bir eksiği olacaktır. eğitimin 3 düzeyi için 2 gölge değişken belirlemeliyiz. Üç eğitim grubunun aynı eğime fakat farklı başlangıç noktalarına sahip olduğunu varsayalım. Bu durumda aşağıdaki modeli yazabiliriz:

$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta X_i + U_i$

Y_i = Yıllık sağlık harcamaları.

X_i = Yıllık gelir.

D_2= 1 Lise eğitimi

= 0 diğer

D_3= 1 Kolej eğitimi

= 0 diğer

Lise eğitim derecesinin altını temel kategori kabul edersek, α_1 bu kategori için başlangıç noktası olur. α_2 ve α_3 diğer kategorilerin temel kategoriden başlangıç noktası farklarını gösterir.

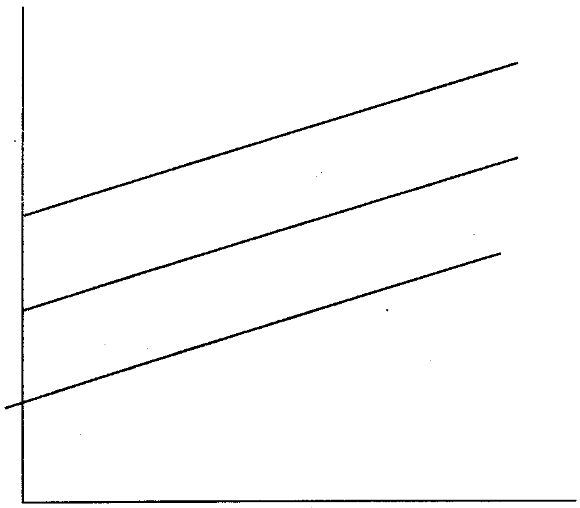
$E(U_i) = 0$ kabul edilirse:

$E(Y_i \mid D_2 = 0, D_3 = 0, X_i) = \alpha_1 + \beta X_i$

$E(Y_i \mid D_2 = 1, D_3 = 0, X_i) = (\alpha_1 + \alpha_2) + \beta X_i$

$E(Y_i \mid D_2 = 0, D_3 = 1, X_i) = (\alpha_1 + \alpha_3) + \beta X_i$

Üsttekiler 3 farklı eğitim düzeyi için sağlık harcamları fonksiyonlarıdır. Bu fonksiyonları geometrik olarak gösterirsek :



Şekil :3

Regresyon çalıştırıldıktan sonra α_2 ile α_3 arasındaki farkın anlamlı olup olmadığı kolayca bulunur. $\alpha_2 = \alpha_3 = 0$ olarak bir hipotez testi, AOV tekniği ve F testi ile aynı anda yapılabilir.

Gölge değişkenlere farklı değerler atayarak regresyonu yorumlayabiliriz. Böylece, eğer $D_2 = 1$

ve $D_3 = 1$ olursa temel kategori kolej eğitimi olur ve bütün karşılaştırmalar bu kategori ile yapılır.

IV-BİR KANTİTATİF VE İKİ KALİTATİF DEĞİŞKEN ÜZERİNDE REGRESYON

Gölge değişken tekniği kolaylıkla birden fazla kalitatif değişken için de uygulanabilir. Kolej öğretmenlerinin geliri için yapılan regresyona geri dönelim. Bu regresyona gelir ve cinsiyete ek olarak öğretmenin rengi de geliri etkileyen başlıca önemli değişken olarak eklendiğini varsayalım. Basit olark renk siyah ve beyaz olarak iki kategoriye ayrılsın.

$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta X_i + U_i$

Y_i = Yıllık gelir.

X_i = Öğretmenlerin yıl olarak deneyimi.

D_2= 1 Erkek

= 0 diğer

D_3= 1 Beyaz

= 0 diğer

Dikkat edilecek nokta cinsiyet ve renk olarak iki kalitatif değişken iki kategoriye sahiptir. Bu yüzden herbiri bir gölge değişkene ihtiyaç duyar. Burada temel kategori siyah kadın öğretmenlerdir.

$E(U_i) = 0$ kabul edilirse:

Siyah kadın öğretmenler için ortalama gelir:

$E(Y_i \mid D_2 = 0, D_3 = 0, X_i) = \alpha_1 + \beta X_i$

Siyah erkek öğretmenler için ortalama gelir:

$E(Y_i \mid D_2 = 1, D_3 = 0, X_i) = (\alpha_1 + \alpha_2) + \beta X_i$

Beyaz kadın öğretmenler için ortalama gelir:

$E(Y_i \mid D_2 = 0, D_3 = 1, X_i) = (\alpha_1 + \alpha_3) + \beta X_i$

Beyaz erkek öğretmenler için ortalama gelir:

$E(Y_i \mid D_2 = 1, D_3 = 1, X_i) = (\alpha_1 + \alpha_2 + \alpha_3) + \beta X_i$

Burada görüldüğü gibi yalnızca başlangıç noktaları katsayıları farklı, eğim katsayısı (β) aynı.

Model üzerinde bir OLS tahmini, hipotezlerin değişik biçimlerini test etmemizi sağlar. Böylece, eğer α_3 anlamlı ise, bu, rengin bir öğretmenin geliri üzerinde etkili olduğu anlamına gelir. Benzer şekilde α_2 anlamlı ise cinsiyetin de bir öğretmenin geliri üzerinde etkili olduğu anlamı çıkar. Eğer bütün başlangıç değerleri anlamlı ise bu, cinsiyetin renk kadar öğretmenlerin geliri üzerinde önemli bir belirleyici değişken olduğu sonucunu verir.

Sonuçta bu model birden fazla kantitatif değişken ve ikiden fazla kalitatif değişken üzerinde uygulanabilir. Unutulmaması gereken her kalitatif değişken için gölge değişken sayısının, o değişkenin kategori sayısından bir eksik olacağıdır. Aşağıda bu durmla ilgili bir örnek verilmektedir.

Örnek:

Bir tanesi birincil diğerleri (bir veya daha fazla olmak üzere) ikincil olarak bir personel iki veya daha fazla işe bakabilir. Bu durum "moonlighter" olarak bilinir. Shisko ve Rostker, moonlighter ücretlerini belirleyen etkenleri bulmaya çalışmışlardır. 318 mmonlighter örnek seçilerek aşağıdaki regresyon modeli elde edilmiştir.

$W_m =$ 37.07 + 0.403 W_0 - 90.06 (Irk) + 75.51 (Kentli) + 47.33 (Eitim) + 113.64 (Bölge) + 2.26 (Yaş)

(.062) (24.47) (21.60) (23.42)
(27.62) (0.94)

$R^2 = 0.34$ df = 311

W_m = Moonlighting ücreti (sent/saat)

W_0 = birincil ücret (sent/saat)

Irk = 0 Beyaz

= 1 Beyaz olmayan

Kentli = 0 Kentli olmayan

= 1 Kentli

Bölge = 0 Batının dışında kalan bölgeler

= 1 Batı

Eğitim = 0 Eğitimi olmayan

= 1 Lise eğitimi olan

Modelde iki kantitatif değişken (W_0 ve yaş) ve dört kalitatif değişken var. Bütün bu değişkenlerin katsayıları %5 derecesinde istatistik olarak anlamlı. Yukarıdaki regresyondan birçok regresyon türetmek mümkündür. Bu regresyonlardan ikisi aşağıda verilmiştir.

Beyaz, kentli olmayan, batılı olmayan, eğitimsiz moonlighter'ların haftalık ücret ortalaması (bütün gölge değişkenler 0)

$W_m = 37.07 + 0.403\ W_0 + 2.26$ Yaş

Beyaz olmayan, kentli , batılı , lise eğitimlilerin haftalık ücret ortalaması (bütün gölge değişkenler 0)

$W_m = 183.49 + 0.403\ W_0 + 2.26$ Yaş

V-İKİ REGRESYONUN KARŞILAŞTIRILMASI

Şimdiye kadar kalitatif değişkenlerin, başlangıç noktasını etkilediğini fakat değişik altgrup regresyonların eğim katsayılarını etkilemediğini varsaydık. Fakat durum böyle değilse ne olacak? Eğer eğimler gerçekte farklıysa. başlangıç noktalarındaki farklılıklar için test yapılabilir. Böylece, iki veya daha fazla regresyonun birbirinden farklı olup olmadığını bulmak için genel bir yönteme gerek duyulur. Farklılık, başlangıç noktalarında veya eğimlerde veya her ikisinde birlikte olabilir. Bu duruma ilişkin bir örnek aşağıda verilmektedir.

Tasarruf ve Gelir, UK 1946 - 1963

Veriler, 1946 - 1954 (ikinci dünya savaşı sonrası, yeniden yapılanma dönemi) ve 1955 - 1963 (yeniden yapılanma sonrası dönemi) olarak ikiye ayrılmıştır. Bulunmaya çalışılan, iki dönem arasında tasarruf-gelir ilişkisinin değişip değişmedigidir.

Yeniden yapılanma dönemi :

$Y_i = \lambda_1 + \lambda_2 X_i + U_{1i}$

$i = 1, 2, \ldots, N_1$

Yeniden yapılanma sonrası dönem :

$Y_i = \gamma_1 + \gamma_2 X_i + U_{2i}$

$i = 1, 2, \ldots, N_2$

Y = Tasarruf

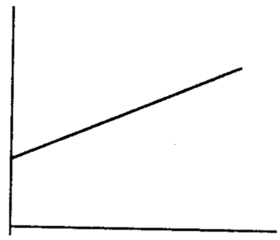
X = Gelir

U_{1i} , U_{2i} = İki regresyondaki hata

Not: N_1 ve N_2 gözlem değerlerinin (iki gruptaki) aynı olmasına gerek yoktur.

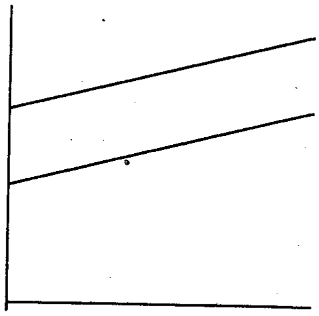
Bu iki regresyon için dört olasılık vardır:

1.) Raslantısal Regresyonlar : $\lambda_1 = \gamma_1$ ve $\lambda_2 = \gamma_2$ iki regresyonun aynı olduğunu gösterir.



Şekil :4

2.) Paralel Regresyonlar : $\lambda_1 \neq \gamma_1$ fakat $\lambda_2 = \gamma_2$ belirli bölümlerinde (başlangıç noktası) regresyonlar farklı.



Şekil :5

3.) Bir Noktada Kesişen Regresyonler : $\lambda_1 = \gamma_1$ fakat $\lambda_2 \neq \gamma_2$ ise her iki regresyonun başlangıç noktaları aynı, fakat eğimleri aynı.

4.) Farklı Regresyonlar : $\lambda_1 \neq \gamma_1$ ve $\lambda_2 \neq \gamma_2$ ise iki regresyon tümden farklıdır.

Verilen değerlerle iki regresyonu çalıştırıp daha sonra bütün olasılıklar için uygun istatistik teknikleri kullanılabilir. Bu tekniklerden biri Chow testi diğeri ise gölge değişkenler tekniğidir.

VI-İKİ REGRESYONUN KARŞILAŞTIRILMASI : CHOW TESTİ

İki veya daha fazla regresyon arasındaki farkları test etmek için yaygın olarak kullanılan yöntemlerden biri Chow testidir. Bu test aşagıdaki kabullerle yapılır:

(a) $U_{1i} \sim N(0,\sigma^2)$

$U_{2i} \sim N(0,\sigma^2)$

(b) U_{1i} ve U_{2i} birbirinden bağımsız dağılıyor.

Tasarruf-gelir regresyonları normal dağılmış olup, ortalaması 0 ve homoskedastik varyans (σ^2) vardır. İki regresyon birbirinden bağımsızdır. Aşagıda tasarruf-gelir verileri ile değişik adımlarda Chow testi yapılmaktadır[3].

Adım 1 : İki dönemdeki N_1 ve N_2 gözlem değerleri birlikte ele alınır ve aşagıdaki tek regresyon çalıştırılır. (Not: Bu örnekte $N_1=N_2=9$ alınmıştır, fakat her zaman aynı değerler olmayabilir.)

$Y_t = \alpha + \beta X_t + U_t$

Bu regresyondan S_1 olarak adlandırdıgımız artık karelerin toplamı (RSS) elde edilir. $df=N_1+N_2-k$, k tahmin edilen parametre sayısı, örneğimizde k=2'dir.

Adım 2 : İki regresyon ayrı ayrı çalıştırılır ve RSS'leri elde edilir. Bunlara S_2 ve S_3 denir.

$df=N_1 - k$ ve $N_2 - k$, k=2. Her iki S'yi toplarsak:$S_4=S_2 + S_3$, $df= N_1+N_2 - 2k$

Adım 3 : $S_5 = S_1 - S_4$ elde edilir.

Adım 4 : F testi uygulanır:

$$F = \frac{S_5/k}{S_4(N_1 + N_2 - 2k)}$$

$df=k, N_1+N_2-2k$. Eğer hesaplanan F değeri kritik F değerini aşıyorsa, iki regresyonun aynı olduğu konusundaki hipotez red edilir.

Bu aşamalar tasarruf-gelir örneğinde uygulanabilir.

Adım 1 : $Y_t = -1.0821 + 0.1178\ X_t$

$(0.1451)\ (0.088)$

$t = (-7.4576)\ (13.3864)$

$r^2 = 0.9185 \quad S_1 = 0.5722 \quad df = 16$

Adım 2 : Yeniden yapılanma dönemi:

$Y_t = -0.266 + 0.0470\ X_t$

$(0.3053)\ (0.0266)$

$t = (-0.8719)\ (1.7669)$

$r^2 = 0.3092 \quad S_2 = 0.1392 \quad df = 7$

Yeniden yapılanma sonrası dönem:

$Y_t = -1.7502 + 0.1504\ X_t$

$(0.3576)\ (0.0175)$

$t = (-4.8943)\ (8.5943)$

$r^2 = 0.9131 \quad S_3 = 0.1931 \quad df = 7$

$S_4 = (S_2 + S_3) = 0.3327$

Adım 3 : $S_5 = S_1 - S_4 = 0.2395$

Adım 4 :

$$F = \frac{0.2395/2}{0.3327/14} = 5.04$$

$F_{2,14}$, %5 için 3.74 'dür. Böylece 5.04 F değeri bu düzeyde anlamlıdır. Sonuç olarak iki regresyon farklıdır. Bu durum aşagıdaki şekilde görülebilir.

tasarruf

Y=-1.75+0.1504X (yeniden yapılanma sonrası dönem)

Y=-0.27+0.0470(yenidenyapılanma dönemi)

Açıkca görüldüğü gibi Chow testi, en küçük kareler yönteminin özel bir uygulamasıdır. İki regresyon yerine tek regresyon tahmin edildiğinde; $\lambda_1 = \gamma_1 = \alpha$ ve $\lambda_2 = \gamma_2 = \beta$ olacaktır.

Chow testinde iki nokta gözönünde bulundurulmalıdır. Birincisi, homokedasite tarafından bozulan tahmin durumunda test uygulanamaz. Bu nedenle ilk olarak homokedasite kontrol edilmelidir. Regresyonu tek başına çalıştırmak için serbestlik derecesinin yeterli olduğunu varsayalım. Uç bir durum olarak: eğer tasrruf-gelir örneğinde ikinci dönem için 2 gözlem değeri alırsak (N-2) = (2-2) = 0 olduğundan serbestlik derecesi 0 olacaktır. Sonuç olarak, $\sigma^2_2 = RSS_2 / df$ belirlenemez. Bu gibi durumlarda :

$$F = \frac{S_5/k}{S_4(N_1 + N_2 - 2k)}$$

olarak F testi değiştirilir. Bu da ikinci gözönünde bulundurlması gereken noktadır.

Yukarıda incelediğimiz örnekte iki altgrup vardı. Chow testi kolaylıkla iki gruptan fazlası için de genelleştirilebilir. Aynı şekilde adım 2'nin dışında dört aşama izlenir. Birçok regresyon bunların altgrupları olabilir ve S_4 ve serbestlik dereceleri elde edilir.

VII-İKİ REGRESYONUN KARŞILAŞTIRILMASI : GÖLGE DEĞİŞKEN YAKLAŞIMI

Çok aşamalı Chow testi işlemi gölge değişkenler kullanılarak kısaltılabilir. Hernekadar Chow testi ve gölge değişken tekniğinden çıkan sonuçlar tümüyle aynıysa da, gölge değişken tekniğinin bazı avantajları vardır. Bu avantajları aşagıdaki örnek üzerinde gösterebiliriz.

Örnek: N_1 ve N_2 gözlemleri var ve aşagıdaki regresyon tahmin ediliyor:

$Y_i = \alpha_1 + \alpha_2 D_i + \beta_1 X_i + \beta_2 (D_i X_i) + U_t$

Y_i = tasarruf

X_i = gelir

D_i = 1 yeniden yapılanma dönemi

= 0 yeniden yapılanma sonrası dönem

$E(U_i)=0$ olduğunu kabul edilerek birinci ve ikinci dönem için tasarruf fonksiyonları elde edilir:

$E(Y_i \mid D_i = 0, X_i) = \alpha_1 + \beta_1 X_i$

$E(Y_i \mid D_i = 1, X_i) = (\alpha_1 + \alpha_2) + (\beta_1 + \beta_2) X_i$

$\gamma_1 = \alpha_1$, $\gamma_2 = \beta_2$, $\lambda_1 = (\alpha_1 + \alpha_2)$ ve $\lambda_2 = (\beta_1 + \beta_2)$

Böylece üstteki tek fonksiyonu tahmin etmekle elde ettiğimiz iki fonksiyonun tahmini birbirine eşittir.

α_2 başlangıç noktası farkı ve β_2 eğim katsayısı farkıdır. β_2, birinci dönemin tasrruf fonksiyonunun eğim katsayısı ile ikinci dönemin tasarruf fonksiyonunun eğim katsayısı arasındaki farkı gösterir. Gölge değişkenlerin toplama (ekleme) formu ise bize iki dönemin başlangıç noktaları arasındaki farkı verir.

$Y_t = -1.7502 + 1.4839 D_i + 0.1504 X_t - 0.1034 X_t$

(0.3319) (0.4704) (0.0163) (0.0332)

t = (-5.2733) (3.1545) (9.2270) (-3.1144)

$R^2 = 0.9425$

Bu regresyon, başlangıç noktaları ve eğim katsayıları arasındaki farkın istatistik olarak anlamlı olduğunu gösteriyor. İki dönem için regresyonların birbirinden farklı olduğu görülüyor.

Yeniden yapılanma dönemi:

$Y_t = (-1.7502 + 1.4839) + (0.1504 - 0.1034) X_t$

$= -0.2663 + 0.0470 X_t$

Yeniden yapılanma sonrası dönem:

$Y_t = -1.7502 + 0.1504 X_t$

Üstteki regresyonlar (görülebilecegi gibi) Chow testi ile elde edilen regresyonların aynısıdır. Gölge değişken tekniğinin Chow testi'nden üstünlüğü kolayca görülebilir. Gölge değişken tekniğinde tek regresyon tahmin edildi, Chow testi ile üç regresyon denklemi ve bunların bileşimi ayrı ayrı tahmin edildi.

1.) Yalnız bir regresyona gerek duyuldu , çünkü tek regresyon denkleminden sonuç çıkarmak daha kolaydır.

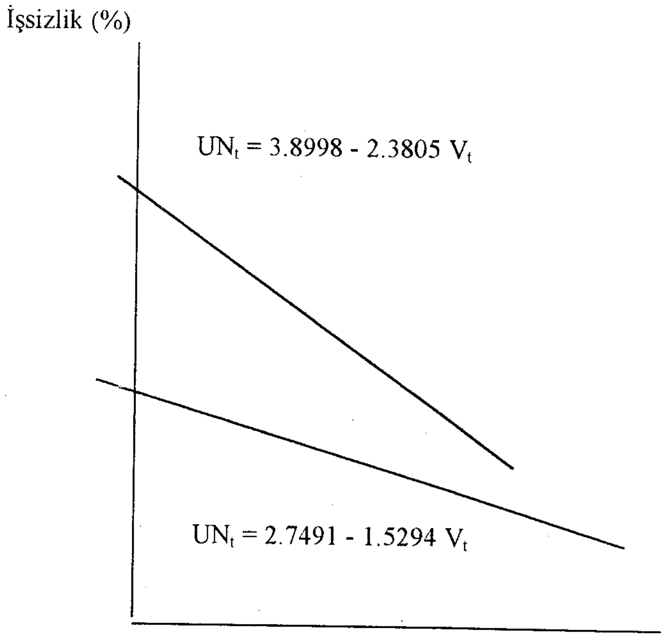
2.) Tek regresyon denklemi değişik hipotez testleri için kullanılabilir. Böylece; başlangıç noktası katsayısı farkı α_2 istatistiksel olarak anlamlıysa iki regresyonun aynı başlangıç noktasına sahip olduğu hipotezi red edilir. Benzer şekilde β_2 8eğim katsayısı farkı) istatistiksel olarak anlamsız fakat α_2 anlamlı ise iki regresyonun aynı eğime sahip olduğu (paralel regresyonlar) hipotezi en azından red edilmez. Bütün regresyonunun kararlılık testi ($\alpha_2=\beta_2=0$) tahmin edilen regresyonun bütünü üstünde F testi uygulanarak yapılır. Eğer bu hipotez sürekli ise regresyon çizgileri üst üste gelecektir.

3.) Chow testi; katsayı, başlangıç noktası veya eğimin farkını açıkca söylemez. Bunlardan bir tanesi Chow testi için anlamlıdır. Bu da yalnızca eğim farkı veya başlangıç noktası farkı veya bütün farklar birlikte olabilir. Bu bakımdan gölge değişken yaklaşımı farklı bir üstünlüğe sahiptir. Gülge değişken yaklaşımı; iki regresyon farklı fakat farklılık kaynakları da (başlangıç noktası, eğim veya ikisi birlikte) farklıysa bunları bize gösterir.

4.) Son olarak , serbestlik derecesinin artışıyla, tahmin edilen parametrelerin göreli kesinliği arttırılabilir.

Örnek:

Bu çalışmada 1958-4 , 1971-2 dönemleri için İngiltere'deki işsizlik oranı ile boş işyeri oranı arasındaki ilişki araştırılmıştır ve aşağıdaki serpilme diyagramı elde edilmiştir. Şekilde görüldüğü gibi işsizlik - boş işyeri ilişkisi 1966'nın 4. çeyreğinden başlayarak değişmektedir. 1966'nın 4. çeyreği ile başlayan işsizlik - boş işyeri ilişkisindeki gözlenen sapmanın istatistiksel olarak anlamlı olup olmadığı aşağıdaki modelle bulunmaya çalışılmıştır.



$Un_t = \alpha_1 + \alpha_2 D_t + \beta_1 V_t + \beta_2 (D_t V_t) + U_t$

UN = İşsizlik oranı (%)

V = Boş işyeri oranı (%)

D= 1 1966-4 ile başlayan dönem

= 0 1966-4'den önceki dönem

t= zaman (bir yılın dörtte biri olarak alınır)

1958-4 ile 1971-2 arasındaki 51 üç aylık gözlem değerleri üzerinde aşağıdaki sonuçlar elde edilmiştir.

$Un_t = 2.7491 + 1.1507\, D_t - 1.5294\, V_t - 0.8511\, (D_t\, V_t)$
(0.1022) (0.3171) (0.1218) (0.4294)
t = (26.896) (3.6288) (-12.5552) (-1.9819)
$R^2 = 0.9128$

Dikkat edilirse başlangıç noktası ve eğim katsayısı farkı %5 düzeyinde istatistik olarak anlamlıdır. Böylece 1966'nın 4. çeyreği ile başlayan UN-V ilişkisindeki gelişimi belirleyen hipotez kabul edilebilir.

Yukarıdaki regresyondan aşağıdaki regresyonlar elde edilebilir.
1958-4 1966-3
$UN_t = 2.7491 - 1.5294\, V_t$
1966-4 1971-2
$UN_t = (2.7491 + 1.15) - (1.5294 + 0.8511)\, V_t - 2.3805\, V_t$
$= 3.8998$

Bu regresyonlar , (üstteki şekilde görülmektedir) 1966-4 ile başlayan dönemdeki UN-V eğrisinin 1958-4 ile başlayan döneme göre çok dik bir eğime ve yüksek bir başlangıç noktasına sahip olduğunu göstermektedir.

VIII-KARŞILIKLI ETKİLEŞİM

Aşağıdaki modeli ele alırsak :
$Y_i = \alpha_1 + \alpha_2\, D_{2i} + \alpha_3\, D_{3i} + \beta\, X_i + U_i$
Y_i = Yıllık giyim harcamaları
X_i = Gelir
D_2 = 1 Kadın
= 0 Erkek
D_3 = 1 Kolej eğitimi olanlar
= 0 Diğer

Bu model kapalı bir model olarak ele alındığında, iki eğitim düzeyi karşısında cinsiyet gölge değişkeni

D_2'nin sabit ve iki cinsiyet düzeyinde eğitim düzeyi gölge değişkeni D_3 'ün sabit olduğu kabul edilir. Bunun anlamı şudur. Kadınlar erkeklerden daha fazla giyim harcaması yapıyorlar fakat bunun eğitimle ilgisi yok. Bunun gibi kolej eğitimi olanlar daha fazla giyim harcamasında bulunuyorlar bunun cinsiyetle ilgisi yok. Birçok uygulamada böylesi bir varsayım savunulamaz. Kolej eğitimli bir kadın, kolej eğitimi görmemiş bir kadından daha fazla giyim harcaması yapabilir. Diğer bir deyişle, iki kantitatif değişken D_2 ve D_3 arasında etkileşim vardır ve bu yüzden bu değişkenler toplam şeklinde değil çarpım şeklinde ortaya çıkarlar.

$Y_i = \alpha_1 + \alpha_2\, D_{2i} + \alpha_3\, D_{3i} + \alpha_4\, (D_{2i}\, D_{3i}) + \beta\, X_i + U_i$
$E(Y_i \mid D_2=1, D_3=1, X_i) = (\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) + \beta\, X_i$
α_2 = Bir kadın için fark
α_3 = Kolej mezunu farkı
α_4 = Kolej mezunu kadını farkı

α_4 . kolej mezunu kadınların giyim harcamasının ortalamasının, kadınların giyim harcaması ortalamasından veya kolej eğitimlilerinin giyim harcamalarından farkını gösterir. Eğer α_2, α_3 ve α_4 pozitif ise kolej mezunu kadınların giyim harcamaları ortalaması temel kategoriye (erkek ve kolej derecesi olmayan) göre daha yüksektir.

Etkileşimli gölge değişkenin katsayısının istatistiksel olarak anlamlı olup olmadığını test etmek için (t) testi kullanılır. Eğer anlamlılık varsa iki özelliğin aynı anda (simultane) varlığı, bu özelliklerin tek tek etkileri artıracak veya azaltacaktır.

IX-MEVSİM ANALİZİNDE GÖLGE DEĞİŞKENLERİN KULLANIMI

Birçok ekonomik zaman serisi aylık veya üç aylık verilerle mevsimsel bir özellik gösterir. Örnek olarak; yılbaşında satışlar, tatilde para talebi, yaz boyunca dondurma ve içecek talebi verilebilir. Genelde yapılan bir zaman serisinden mevsimsel faktörü çıkararak trend gibi diğer faktörler üzerinde yoğunlaşmaktır. Zaman serisinden mevsim faktörünü çıkarma işlemine mevsim düzeltmesi adı verilir ve böylesi bir zaman serisine de mevsim düzeltmeli zaman serisi denir. Önemli ekonomik zaman serileri (fiyat indeksleri, endüstriyel üretim indeksleri) genellikle mevsim düzeltmesi yapılarak yayınlanır.

Bir zaman serisi üzerinde mevsim düzeltmesi için bir çok yöntem olmasına karşın biz bu yöntemlerden yalnızca birini; gölge değişken yöntemin ele alacağız. gölge değişken yöntemi ile nasıl mevsim düzeltmesi yapıldığını göstermek için US üretim şirketinin karını 1965 - 1970 dönemlerinin 3 aylık satış verileri üzerinde tahmin edebiliriz.

$Kar_t = \alpha_1 + \alpha_2\, D_{2t} + \alpha_3\, D_{3t} + \alpha_4\, D_{4t} + \beta\, (\text{Satışlar})_t + U_t$
D_2 = 1 İkinci 3 aylık dönem
= 0 Diğer
D_3 = 1 Üçüncü 3 aylık dönem
= 0 Diğer
D_4 = 1 Dördüncü 3 aylık dönem
= 0 Diğer

Mevsim değişkeni 4 sınıfa ayrılır. Bir yıl 4 tane üçer aylık döneme ayrılır, böylece 3 tane gölge değişkene gerek duyulur. Böylece, eğer değişik 3 aylık dönemlerde mevsimsel bir model varsa (en iyi temsil eden model) α_2, α_3 ve α_4 tahmin edilebilir. model genelde bütün durumlara uyum içindedir. (Not: birinci 3 aylık dönem temel kategoridir.)

$Kar_t = 6688.3789 + 1322.8938\, D_{2t} - 217.8037\, D_{3t} + 183.8597\, D_{4t} + 0.0383\, (Satışlar)_t$
(1711.3707) (638.4753) (632.2561)
(654.2937) (0.0115)
t = (3.9082) (2.0720)(-0.3445) (0.2810) (3.3313)
$R^2 = 0.5255$

Bu sonuçlar bize satış katsayısı ve ikinci 3 aylık dönemle birlikte başlangıç noktası farkının %5 güven sınırında istatistiksel olarak anlamlı olduğunu gösterir. Her yılın ikinci döneminde işleme giren bazı mevsimsel faktörler olduğu sonucuna böylece varılabilir. 0.0383 olan satış katsayısı , satış 1 $ arttırılırsa kar 4 sent artar sonucunu verir. Temel veya birinci 3 aylık dönem için karın ortalama düzeyi 6688 $ ve ikinci 3 aylık dönem için de 1323 $ fazla olarak 8011 $ 'dır.

İkinci 3 aylık dönemi diğer dönemlerden ayırmak için tek gölge değişken kullanıldığında:

$Y_t = 6515.6 + 1331.4\ D_2 + 0.0393\ (Satış)$
(1623.1) (493.02) (0.0106)
t= (4.0143) (2.7004) (3.7173)
$R^2 = 0.5155$
$D_2 = 1$ İkinci 3 aylık dönem
$= 0$ Diğer

Üstteki iki regresyonda birinci, üçüncü ve dördüncü dönemlerin başlangıç noktaları eşit kabul edilmiştir. Böylesi bir ayrım daha önce görüldüğü gibi geçerlidir.
İkinci 3 aylık dönem (Kar=8011.2727+0.0383 (Satış))
Birinci 3 aylık dönem (Kar=6688.3789+0.0383 (Satış))

İlk modelde yalnızca 3 aylık dönemlerdeki başlangıç terimleri farklı fakat satış değişkeninin eğim katsayısı her 3 aylık dönemde aynıdır. Ancak bu kabuller çarpımlı gölge değişken tekniği ile test edilmelidir.

X-PARÇALI DOĞRUSAL REGRESYON

Aşağıdaki şekilde gölge değişkenlerin bir başka kullanım yeri olarak bir işletmenin satışı ödüllendirme (prim verme) durumu gösterilmektedir. Bir eşik değeri üzerinde satış üzerinden komisyon ödenir. Bu eşik değeri X^* düzeyi olarak gösterilir. Stokastik bir komisyon yapısı ve arkasında bir başka düzey vardır. (Not: Satış komisyonunu etkileyen diğer faktörler stokastik karışıklık ifadesi ile gösterilir.) satış komisyonunun X^* düzeyine kadar doğrusal arttığı varsayılmıştır. Satış komisyonu, X^* düzeyinden sonra da doğrusal fakat daha hızlı artmaktadır. Böylece iki parçadan oluşan şekil üzerinde 1 ve 2 olarak gösterilebilinecek bir "parçalı doğrusal regresyon" vardır ve komisyon fonksiyonunun eğimi eşik değerinde değişmektedir. Komisyon, satış verileri ve X^* düzeyinin değeri verildiğinde, gölge değişken tekniği , şekilde gösterilen parçalı doğrusal regresyonun iki parçasının eğimlerini tahminde kullanılabilir.

$Y_i = \alpha_1 + \beta_1 X_i + \beta_2 (X_i - X^*) D_i + U_i$
Y_i = Satış komisyonu
X_i = Satış elemanının sattığı miktar
X^* = Satış eşik değeri
$D = 1 \quad X_i > X^*$
$= 0 \quad X_i < X^*$
$E(U_i) = 0$ olduğu varsayılarak:
$E(Y_i \mid D_i = 0, X_i, X^*) = \alpha_1 + \beta_1 X_i$
$E(Y_i \mid D_i = 1, X_i, X^*) = \alpha_1 - \beta_2 X^* + (\beta_1 + \beta_2) Xi$

Böylece β_1 , birinci bölümdeki regresyonun eğimini $(\beta_1 + \beta_2)$ ise ikinci bölümdeki regresyonun eğimini verir.

XI-SONUÇ

Bu çalışmanın amacı kalitatif veya gölge değişkenlerin regresyon modellerinde kantitatif değişkenlerle birlikte nasıl kullanıldığını göstermektir. Gölge değişkenler gerçekte bir veri sınıflama aracı olarak verileri kalitatif veya nitelik (cinsiyet , din, bölge...v.b.) temelinde değişik altgruplara ayırırlar ve her altgrup için regresyon modeli oluşturarak bu modeli çalıştırırlar. Eğer değişik altgruplardaki kantitatif değişkenlerin varyansı farklıysa bu farklılık, değişik altgruplardaki regresyonların başlangıç noktası veya eğim katsayısı veya her ikisinin birden farklı olmasına yol açacaktır.

Her nekadar gölge değişken tekniği çok yönlü ise de dikkatli kullanmayı gerektirir. Birincisi, eğer regresyon modeli sabit bir terim içeriyorsa gölge değişkenlerin sayısı, sınıflanmış kalitatif değişken sayısının bir eksiği olmalıdır. İkinci olarak; gölge değişkenin katsayısı, temel grupla (baz alınan grup) ilişkilendirilerek yorumlanmak zorundadır. Son olarak; eğer bir model birçok sınıfla birlikte birçok kalitatif deişkene sahipse, gölge değişken için serbestlik derecesi yüksek olmalıdır.

KAYNAKLAR

[1]-Damor Gujarati; Basic Ekonometrics; Singapour,1994.

[2]-Parsaye Kamran - Chignell Mark- Khoshafian Setrag-Wong Harry; Intelligent Databases (New York, Wiley,1989)

[3]-Ullman Jeffrey D.; A Comparison Between Deductive And Object Oriented Database Systems (press unknown).