3. ve 4. Hafta

HETEROSKEDASİTE (DEĞİŞEN VARYANS)

Klasik doğrusal regresyon modelinin temel varsayımlarından biri, ana kütle regresyon fonksiyonundaki rassal hata terimlerinin (u_i) sabit varyanslı, diğer bir ifade ile her alt ana kütle için hata teriminin varyansının eşit olmasıdır.

$$Var(u_i) = E\left[u_i - E(u_i)\right]^2 = E(u_i^2) = \sigma^2$$

Hata terimi varyansının bütün gözlemler için aynı olması ***homoskedasite (sabit varyans),*** hata teriminin varyansının gözlemler itibariyle değiştiği durum ise ***heteroskedasite (değişen varyans)*** olarak adlandırılmaktadır.

Homoskedasite varsayımına göre; X_i'lere bağlı Y'nin koşullu varyansı X hangi değeri alırsa alsın sabittir ve hata teriminin koşullu varyansı da Y'nin koşullu varyansına eşittir. Bilindiği üzere Y'nin koşullu varyansı, gözlemlenen Y_i değerlerinin Y'nin beklenen değeri $E(Y|X = X_i)$ etrafındaki dağılımını gösterir. Heteroskedasite ddurumunda ise X değişkeninin alacağı değere göre hata teriminin varyansıda değişmektedir. Örneğin X'in değeri arttıkça Y'nin koşullu varyansı büyür veya X'in değeri arttıkça Y'nin koşullu varyansı küçülür.

$$E(u_i^2) = \sigma_1^2 = h(X_i)$$

Hane halkı tüketim harcaması ile haftalık gelirin yer aldığı ana kütle için, hane halkının haftalık tüketim için gerçekleştirdiği ortalama tüketim $E(Y_i)$, hane halkı gelirinin (X) doğrusal bir fonksiyon olarak tanımlanmıştı.

$$E(Y) = \beta_0 + \beta_1 X$$

Bilinmeyen ana kütle parametreleri (β_0 ve β_1), yukarıda verilen tüketim fonksiyonu hakkında bilgi taşımaktadır. Tepki parametresi olan β_1, hane halkı geliri bir birim arttığında hane halkının ortalama tüketim harcamasının ne kadar değişeceğini gösterir. Sabit terim β_0, gelir düzeyi sıfır iken, gıda harcamalarını ölçer.

β_0 ve β_1'yi tahmin etmek için i=1,2,...,10 şeklinde endekslenen, i. hane halkı için tüketim harcamasını ve gelir düzeyini gösteren (Y_i, X_i) çifti ile n=5 hanelik bir örneklem düşünelim.

Belirli bir gelir düzeyi ile bütün hane halkları aynı tüketim harcamasına sahip olmayacak ve regresyon modeliminin genel tanımlaması doğrultusunda u_i', i. hane halkının tüketim harcaması Y_i ile X_i gelir düzeyinde tüm hane halkının ortalama tüketim harcaması arasındaki fark olarak tanımlanmaktadır.

$$u_i = Y_i - E\left(Y_i\right) = Y_i - \beta_0 - \beta_1 X$$

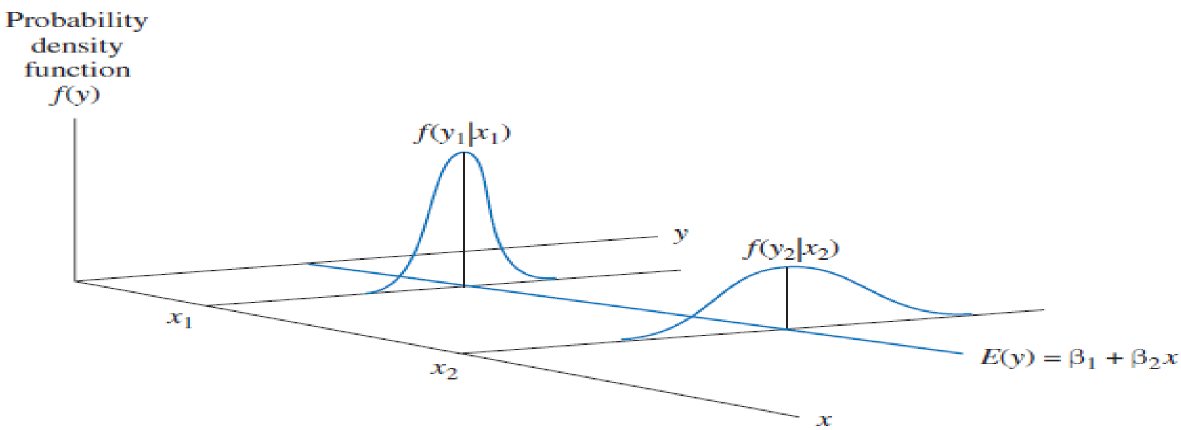
Böylece i. hane halkının tüketim harcamalarını tanımlamak üzere kullanılan model aşağıdaki şekilde yazılır.

$$Y_i = \beta_0 + \beta_1 X + u_i$$

$E(Y) = \beta_0 + \beta_1 X_i$ tüketim harcamalarının gelir düzeyiyle açıklanan bölümünü , u_i yi ise tüketim harcamalarının diğer faktörler tarafından açıklanan bölümünü göstermektedir. Bu bölümde $E(Y) = \beta_0 + \beta_1 X_i$ ortalama tüketim fonksiyonunun düşük gelirli hanelerin tüketim harcamalarını yüksek gelirli hanelerin tüketim harcamalarına göre daha iyi açıklayıp açıklayamadığını sorgulamaya başlıyoruz. Düşük gelirli hanelerin tüketim harcamaları ile yüksek gelirli hanelerin tüketim harcamalarını tahmin etmek zorunda olsaydınız, hangi tahminin daha kolay olacağını düşünürdünüz? Düşük gelirli hane halkları tüketim için fazla seçeneğine sahip değildir, ancak gıda, barınma gibi zorunlu tüketim harcamalarında bulunurlar ve bu tür harcamalar için gelirlerinin büyük bir kısmını kullanırlar. Diğer taraftan, yüksek gelirli hane halklarının tüketim harcamaları çeşitlilik arz eder. Düşük gelirli hane halkları zorunlu harcamaları yapıyor iken, yüksek gelirli aileler yılda bir iki kez tatile çıkabilir, haftanın belli günlerinde dışarıda yemek yiyebilir, spor merkezine gidebilir. Böylece gelir değişkeni nispi olarak yüksek gelirli hane halklarının tüketim harcamalarını açıklamak için daha az önemli bir değişkendir. Yüksek gelirlilerin tüketim harcamalarını tahmin etmek zordur.

Tam olarak ifade ettiğimiz şeyi tanımlamanın bir başka yolu, u_i'nin büyük pozitif veya negatif değerler alma olasılığı, yüksek gelirliler için düşük gelirlilere göre daha yüksek olduğunu söylemektir. Hanenin geliri yüksek ise, gelir dışındaki faktörler tüketim harcamaları üzerinde büyük bir etkiye sahip olabilir.

İktisadi bir gerçek olan yukarıdaki olgu nasıl modellenecektir? Rassal değişken u_i'nin varyansı yüksekse, daha yüksek değerler alma olasılığına sahiptir. Bu etki, $Var(u_i)$ yi doğrudan X'e bağlı bir biçimde tanımlayarak tespit etmek mümkündür. Diğer bir ifade ile, X arttıkça $Var(Y_i)$ artacağını söylemek mümkündür. X'in büyük olduğu gözlemlerde tüketim harcaması bağımlı değişken Y'nin ortalamasından ($E(Y) = \beta_0 + \beta_1 X_i$) daha çok sapabilir. Bu durumda tüm gözlemlerin varyansları aynı olmadığından heteroskedasitenin (değişen varyans) varlığından söz edebiliriz. Alternatif olarak, rassal değişken Y ve rassal hata u_i değişen varyanslıdır. Tersine eğer tüm gözlemler aynı varyansa sahip olasılık yoğunluk fonksiyonundan geliyorsa homoskedasitenin (sabit varyans) söz konusu olduğunu söyleyebiliriz ve Y ve u_i sabit varyanslıdır ki bu durum doğrusal regresyon modelinin temel varsayımlarındandır.



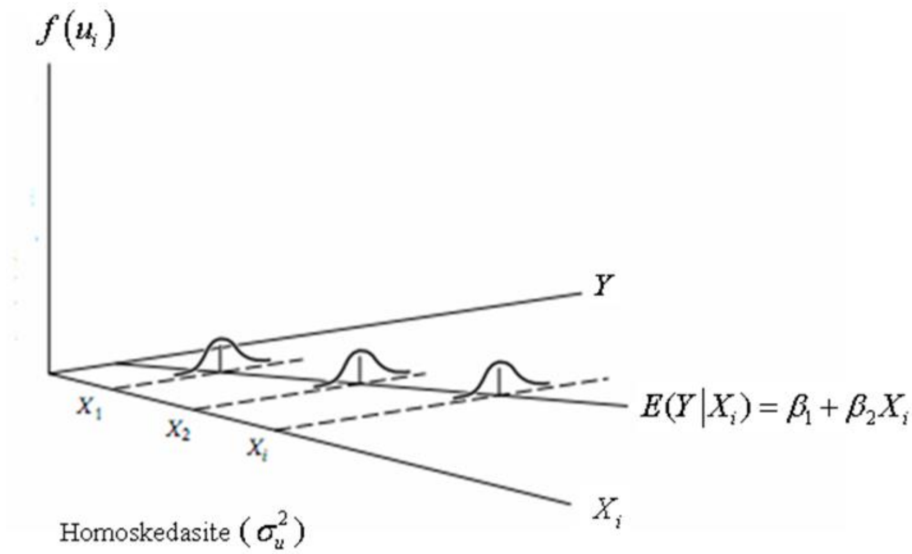
Değişen varyanslı hatalar

Sabit varyans varsayımının yerine gelmediği durum değişen varyans, yukarıdaki şekilde gösterilmiştir. Değişen varyansı yukarıda verilen tüketim harcamaları ve gelir verilerinin yer aldığı ana kütleye ait tablonun son satırından da görmeniz mümkündür.

$P(Y/X)$ \ X_i	80	100	120	140	160	180	200	220	240	260
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
Koşullu Olasılıklar	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	-	1/6	-	1/7	1/6	1/6	-	1/7	1/6	1/7
	-	-	-	1/7	-	-	-	1/7	-	1/7
Y'nin koşullu ortalamaları	65	77	89	101	113	125	137	149	161	173
Y'nin koşullu varyansları= σ^2	50	66	46.4	133.42	57.3	116.6	82.4	112	310.7	216.5

Şöyle ki; gelir arttıkça rassal hatanın varyansı ve dolayısıyla rassal hata ile aynı varyansa sahip olan *Y*'nin varyansı da artmaktadır. Yukarıdaki şekilde $X = X_i$ iken, $f(Y_i, X_i)$ şeklindeki olasılık yoğunluk fonksiyonu, Y_1'in yüksek olasılıkla $E(Y_1)$'e yaklaşacağını ifade etmektedir. X_2'ye doğru gidildikçe olasılık yoğunluk fonksiyonu $f(Y_2, X_2)$ daha hızlı yayılacaktır. Örneğin yukarıdaki ana kütle için $X = 80$ iken rassal hatanın varyansı 50 , $X = 100$ iken rassal hatanın varyansı 66 dır. Buna göre gelir arttıkça rassal hatanın varyansı da artmakta, dolayısıyla varsayımda durumu değişen varyans söz konusudur.

Ancak Y_2'nin düştüğü ve daha büyük değerler aldığı yer hakkında kesin bilgiye sahip değiliz. Sabit varyans söz konusu olduğunda hatalar için olasılık yoğunluk fonksiyonu aşağıdaki şekildeki gibi *X*'deki değişim ile değişmez.



Değişen varyansın varlığının yukarıda da ifade edildiği üzere, küçük kareler varsayımlarının birinin ihlalidir. $Y_i = \beta_0 + \beta_1 X + u_i$ modeli için u_i nin sıfır ortalama, sabit varyanslı (σ^2) ve rassal hata terimlerinin ilişkisiz olduğunu varsayımları kısaca aşağıdaki gibi gösterilmiştir.

$E(u_i) = 0$, $Var(u_i) = \sigma^2$, $Cov(u_i, u_j) = 0 \quad i \neq j$ için

Şimdi yukarıdaki varsayımlardan $Var(Y_i) = Var(u_i) = \sigma^2$ yi ifade eden sabit varyans varsayımı geçerli değildir. Dolayısıyla sabit varyans varsayımının aşağıdaki gibi başka bir varsayım biçimi ile değiştirilmesi gerekmektedir.

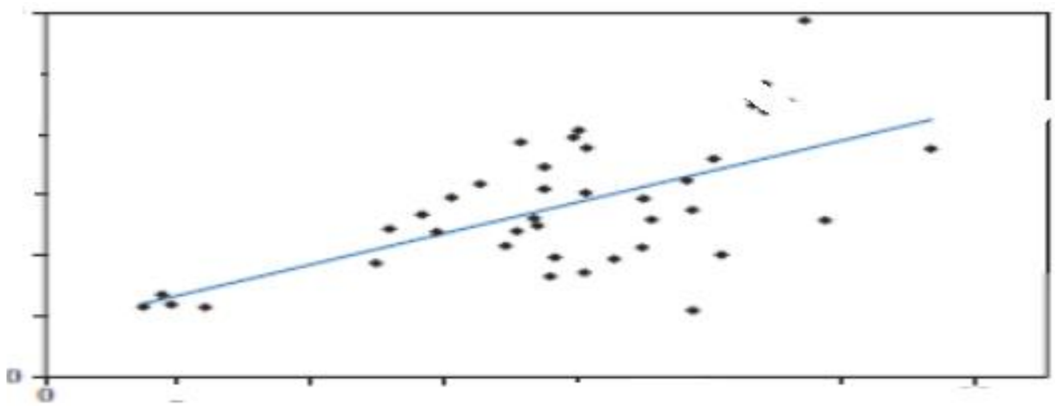
$$Var(Y_i) = Var(u_i) = h(X_i)$$

Burada $h(X_i)$, X_i 'in fonksiyonu olduğu için örneğin X_i arttığında $h(X_i)$ da artacaktır. Hata teriminin heteroskedastik yapıya sahip olduğu modellerde rassal hatanın normal dağılıma sahip olduğu varsayılır.

Değişen varyansın yapısını daha ayrıntılı bir biçimde sunabiliriz ve aynı zamanda ortalama fonksiyonunun ($E(Y) = \beta_0 + \beta_1 X_i$) en küçük kareler tahmini ve ilgili en küçük kareler kalıntılarını yeniden inceleyerek değişen varyansı belirlemek için biçimsel olmayan bir yol sunabiliriz. Tüketim harcamaları modeli Örnek 1 gözlemleriyle tahmin edilen en küçük kareler tahmini aşağıdaki gibidir.

$$\hat{Y}_i = 17.29 + 0.61 X_i$$

Bu tahmini fonksiyonun bir grafiği, gözlemlenen tüm harcama-gelir noktalarıyla birlikte (Y_i, X_i) aşağıdaki şekilde gösterilmiştir. Gelir (X) arttığında, tahmini ortalama fonksiyonundan daha fazla sapan veri noktalarının yaygınlığı artar. X arttıkça regresyon doğrusundan daha uzakta dağılmış noktaların sayısı artar. Bu özelliği tanımlamanın bir başka yolu, aşağıda tanımlandığı gibi, en küçük kareler kalıntılarında gelir arttığında mutlak değer olarak artış eğilimi söz konusu olduğunu söylemektir. $\hat{u}_i = \hat{Y}_i - 17,29 - 0,61X_i$



Şekil: Tüketim harcamaları fonksiyonunun EKK tahmini ve gözlemlenen veri noktaları

Gözlemlenebilen EKK kalıntıları (u_i), $u_i = Y_i - \beta_0 - \beta_1 X_i$ ile verilen gözlemlenemeyen hataların (u_i) tahmini olduğu için yukarıdaki şekil gelir arttığında gözlemlenemeyen hataların da mutlak değer olarak artış eğilimi sergilediğini gösterir. Diğer bir ifade ile gelir (X) arttığında tüketim harcamasının (Y) ortalama tüketim harcaması ($E(Y)$) etrafındaki değişimi artar. Bu tespit, daha önce belirttiğimiz ortalama tüketim harcamaları fonksiyonunun düşük gelirli hane halkları için tüketim harcamalarını açıklamada yüksek gelirli hane halklarından daha iyi olduğunu ifade eden hipotez ile devam eder. Değişen varyans ile Y'nin, ortalaması etrafındaki artan değişimini tespit edebiliriz.

Hata teriminin varyansının gözlemler itibariyle değişmesinin ana sebepleri aykırı (outliner) gözlemler ve model kurma hatasıdır.

Değişen varyans ile daha çok yatay kesit veri kullanıldığında karşılaşılır. Bilindiği üzere yatay kesit veri terimi, zamanın belirli bir noktasında firma ya da hane halkları gibi ekonomik birimlere ait veri setini ifade eder. Gelir ve tüketim harcaması için hane halkı verisi de bir yatay kesit veri setidir. Yatay kesit veri her zaman değişen büyüklükteki ekonomik birimlerin gözlemlerini içerir. Örneğin hane halkı verisi farklı sayıda hane halkı üyelerini ve farklı hane halkı gelir düzeylerini içerecektir. Hane halkının daha yüksek gelir düzeyi ne kadar fazla ise açıklayıcı değişkenler setindeki değişim ile bazı sonuç değişkenlerindeki (*Y*) değişimi açıklamak o kadar zordur. *Y* değerlerinin tespit edilmesi konusunda daha yüksek gelirli hane halklarının daha farklı ve esnek olması muhtemeldir, kısaca homojen değildir. Doğrusal regresyon modelleri için bunun anlamı şudur: Ekonomik birimin boyutunun daha büyük olması sonuç *Y* ile ilişkili daha çok belirsizliğe neden olur. Bu büyük belirsizlik, ekonomik birimin boyutu ne kadar büyükse o kadar büyük olan bir hata varyansı belirlenerek modellenebilir.

Değişen varyans yatay kesit veri ile sınırlandırılması gereken bir özellik değildir. Bir firma, bir hane halkı gibi ekonomik birimlerin veya tüm ekonomiye ilişkin zaman serisi verisi ile hata varyansının değişmesi mümkündür. Eğer bir dışsal şok veya Y hakkında daha fazla ya da daha az belirsizlik yaratacak durumlarda değişimler söz konusu ise bu doğru olacaktır.

En Küçük Kareler Tahmincileri için Değişen Varyansın Sonuçları

Değişen varyans varlığı $Var(u_i)=\sigma^2$ şeklindeki en küçük kareler varsayımının ihlal edildiği anlamına geldiği için bu ihlal, EKK tahmincilerinin özellikleri açısından ne gibi sonuçlar yaratır.

Heteroskedasite durumu, tahmin edilen parametrelerin sahip olması istenen özelliklerden eğilimsizlik (sapmasızlık, sistematik hatasızlık) özelliğini etkilemez. Kısaca heteroskesitenin varlığı durumunda EKK tahmincileri eğilimsizlik özelliğini korurlar.

$$Y_i=\beta_0+\beta_1 X_i+u_i \qquad E(Y|X=X_i)$$

$$\bar{Y}=\beta_0+\beta_1\bar{X}+\bar{u}$$

$$\beta_0=\bar{Y}-\beta_1\bar{X}-\bar{u} \qquad \hat{\beta}_0=\bar{Y}-\hat{\beta}_1\bar{X}$$

$$\hat{\beta}_0-\beta_0=(\bar{Y}-\hat{\beta}_1\bar{X})-(\bar{Y}-\beta_1\bar{X}-\bar{u})$$

$$\hat{\beta}_0-\beta_0=\bar{Y}-\hat{\beta}_1\bar{X}-\bar{Y}+\beta_1\bar{X}+\bar{u}$$

$$\hat{\beta}_0-\beta_0=-\hat{\beta}_1\bar{X}+\beta_1\bar{X}+\bar{u}=-(\hat{\beta}_1-\beta_1)\bar{X}+\bar{u}$$

$$Y_i-\bar{Y}=(\beta_0+\beta_1 X_i+u_i)-(\beta_0+\beta_1\bar{X}+\bar{u})$$

$$Y_i-\bar{Y}=\beta_1(X_i-\bar{X})+(u_i-\bar{u})$$

$$y_i=\beta_1 x_i+u_i^* \qquad \hat{\beta}_1=\sum x_i y_i \Big/ \sum x_i^2$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i(\beta_1 x_i + u_i^*)}{\sum x_i^2}$$

$$\hat{\beta}_1 = \frac{\beta_1 \sum x_i^2}{\sum x_i^2} + \frac{\sum x_i u_i^*}{\sum x_i^2} \qquad \sum x_i u_i^* = 0$$

$$\hat{\beta}_1 = \beta_1$$

Her iki tarafın beklenen değeri alınır.

$$E(\hat{\beta}_1) = \beta_1$$

$$E(\hat{\beta}_0) = \beta_0$$

Eğilimsizlik özelliği hata terimlerinin homoskedasite koşuluna bağlı değildir.Heteroskedasite parametrelerin kararlı ve tutarlı olmalarını elgellememektedir.

Tahmin edilen parametrelerde aranan diğer bir özellik, tahmin edilen parametrelerin en düşük varyansa sahip olması diğer bir ifade ile tesirli (en iyi) olmasıdır. Hata terimlerinn varyansı heteroskedasik ise, EKK yöntemi ile tahmin edilen parametreler en küçük varyansa sahip olamayacaklar, böylelikle tahmin edilen parametreler güvenilir olmayacaktır. Çünkü EKK yöntemi en büyük varyansa ağırlık verecektir. Değişen varyans söz konusu olmaksızın basit bir doğrusal regresyon modeli aşağıdaki gibidir.

$$Y_i = \beta_0 + \beta_1 X + u_i \quad \text{ve} \quad Var(u_i) = \sigma^2$$

β_1'in EKK tahmincisi $\hat{\beta}_1$'in varyansını daha önce aşağıdaki gibi tanımladığımızı hatırlayalım.

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum\left(X_i - \bar{X}\right)^2} = \frac{\sigma^2}{\sum x_i^2}$$

Yukarıdaki tanım rassal hatanın varyansının sabit ve σ^2'ya eşit olduğu varsayımına dayanmakta olup, rassal hatanın varyansının tahmincisi ise $\hat{\sigma}^2 = \sum \hat{u}_i^2 / n-k$ ‘dır. Şimdi her bir gözlemin hata varyanslarının farklı olduğunu (Heteroskedasite) varsayalım ve bu farkı σ^2ye *i* simgesi ekleyerek σ_i^2 ile gösterelim, böylece en küçük kareler tahmincisi $\hat{\beta}_1$'nin varyansı aşağıdaki gibidir.

$$Var(\hat{\beta}_1) = \sum_{i=1}^{n} k_i^2 \sigma_i^2 \qquad k_i = \frac{(X_i - \bar{X})}{\sum (X_i - \bar{X})^2} = \frac{x_i}{\sum x_i^2}$$

$$= \sum_{i=1}^{n} \left[\frac{(X_i - \bar{X})}{\sum (X_i - \bar{X})^2} \right]^2 \sigma_i^2$$

$$= \frac{\sum (X_i - \bar{X})^2}{\left[\sum (X_i - \bar{X})^2 \right]^2} \sigma_i^2$$

$$= \frac{\sum x_i^2 \sigma_i^2}{(\sum x_i^2)^2}$$

Sonuç olarak eğer $Var(u_i) = \sigma^2$ geçerli değil ise, yukarıdaki varyans denkleminde sadeleştirmeler yapılamayacak (toplama işlemcisinin dışına alınamayacak) ve sonuç itibariyle artık

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum \left(X_i - \bar{X} \right)^2} = \frac{\sigma^2}{\sum x_i^2}$$

denklemi geçerli olmayacağı için, varyansın kare kökü standart hatalar da geçerli değildir. Buna bağlı olarak ana kütle parametreleri için oluşturulan güven aralıkları ve hipotez testleri de gücünü ve güvenirliliğini yitirmektedir. Heteroskedastik bir modele EKKY uygulandığında *t* ve*F* testleri doru olmayan sonuçlar verecektir. Çünkü standart hatalar olduğundan büyük değerli elde edilmektedir.

Heteroskedasitenin Tespit Edilmesi

Heteroskedasitenin tespiti için kullanılan testler, sistematik ve sistematik olmayan testlerdir. Sistematik olmayan testler kalıntı grafiklerinin incelenmesidir ki; bu yolla başarı biraz da araştırmacının tecrübesine bağlıdır. En küçük kareler hatalarının grafiklerinin incelenmesi heteroskedasiteyi belirlemede biçimsel olmayan bir yoldur. Sistematik testler ise parametrik ve parametrik olmayan testler olarak da ele alınabilir. Parametrik testler yardımcı bir regresyonun tahminini gerektirirken, parametrik olmayan testler bir test istatistiğine dayanır, yardımcı regresyon tahmin edilmez.

Sistematik olmayan test: Kalıntı Grafikleri

Heteroskedasitenin varlığını araştırmanın bir yolu, modeli en küçük kareler yaklaşımıyla tahmin etmek ve en küçük kareler kalıntılarının grafiğini çizmektir. Eğer hatalar sabit varyanslı ise kalıntılar herhangi bir türden bir örüntü izlemez. Eğer hatalar değişen varyans içeriyorsa onlar sistematik olarak daha büyük değişimler sergilemeye eğilimli olabilir.

Örneğin hane halkı tüketim harcamaları veri seti için gelir arttığında varyansın da artacağından şüphelendik. En küçük karelerin kalıntılarının gelir düzeylerine karşı yukarıdaki grafiğini çizdik. Grafikten gelir arttığında kalıntıların mutlak büyüklüğünün önemli düzeyde arttığı açık olarak görülmektedir.

Değişen varyansı araştıran bu yönteme herhangi bir basit regresyon denklemi için başvurulabilir. Çok değişkenli regresyonda kalıntıların sistematik olarak değişip değişmediğini gözlemlemek için kalıntıların her bir açıklayıcı değişkene ya da $\hat{Y}_i$'ye karşı grafiği çizilerek değişen varyansı görsel olarak tespit etmek mümkündür.

Sistematik Testler

Heteroskedasitenin teorik biçimi:

$$\sigma_i^2 = \sigma^2 X^{\rho}$$

burada ρ heteroskedasite parametresidir. Temel ve alternatif hipotezler ise aşağıdaki gibidir.

$H_0 : \rho = 0$ Homoskedasite

$H_1 : \rho \neq 0$ Heteroskedasite

$\rho = 0$ ise $\sigma_i^2 = \sigma^2 X^{\rho} = \sigma_i^2 = \sigma^2 X^0$ olacaktır. $X^0 = 1$ den $\sigma_i^2 = \sigma^2$ eşitliği elde edilir ki; bu da her gözlemin hata termin varyansının σ^2 gibi sabite eşit olduğunu göstermektedir. Böylece bağımsız değişken hata teriminin varyansının büyüklüğüne etki etmeyecektir.

Spearman Sıra Korelasyon Testi

Spearman Sıra Korelasyon testi parametrik olmayan bir testtir Öncelikle istatistikte kullanılan Sıra korelasyon testinin ne olduğuna kısaca değinelim. *X* ve *Y* değerleri arasındaki korelasyonu hesaplamak yerine, 2 serideki kıymetler büyükten küçüğe ya da küçükten büyüğe doğru numaralandırılarak, bu sıra numaraları arasındaki ilişki araştırılırsa buna sıra korelasyon testi adı verilmektedir. Ekonometride değişen varyansın tespiti için bu testin uygulamasında kalıntı ile X bağımsız değişkenine sıra numarası verilir.

Spearman sıra korelasyon testinin uygulaması aşağıdaki gibidir.

Öncelikle değişen varyansın varlığının araştırıldığı tahmin edilmiş regresyon denkleminden kalıntılar elde edilir. Kalıntılara işaretleri dikkate alınmadan (mutlak değerleri ile) büyükten küçüğe ya da küçükten büyüğe doğru sıra numarası verilir. Aynı işlem *X* bağımsız değişkeni için de yapılır. Her iki sıra numarası arasındaki fark (*d*) hesaplanır. Daha sonra kalıntıların sıra kıymetlerı ile *X* değişkenin sıra kıymetleri arasındaki ilişkiyi gösteren aşağıdaki Spearman sıra korelasyon katsayısı hesaplanır.

$$r_s = 1 - 6\left[\frac{\sum d_i^2}{n\left(n^2 - 1\right)}\right]$$

Buradaki $d_i = X_i - \hat{u}_i$ şeklinde hesaplanır ve sıralama kıymetleri arasındaki farkı ifade eder. *n* gözlem sayısıdır. Temel ve alternatif hipotezler

$$H_0 : \rho_s = 0$$

$$H_1 : \rho_s \neq 0$$

Temel hipotezin doğru olduğu varsayımı altında test istatistiği aşağıdaki gibidir.

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \sim t_{(n-2)}$$

Anakütle korelasyon katsayısının 0’a eşit ve gözlem sayısının 8’den büyük olduğu varsayımına göre, spearman sıra korelasyon testi *n*-2 serbestlik derecesi ile *t* –dağılımına uygunluk göstermektedir

Test istatistiğinin hesaplanan değeri, tablo değeri ile karşılaştırılır. Test istatistiğinin hesaplanan değeri erilen α anlamlılık düzeyinde $t_{(n-2)}$ değerinden büyükse, temel hipotez reddedilir ve alternatif kabul edilir. Bu durumda ρ_s sıfırdan farklı ve istatistiksel açıdan anlamlıdır. Bu Alternatif hipotezin kabulü, *X* ile kalıntılar arasında kuvvetli bir ilişkinin olması yani heteroskedasitenin var olması anlamına gelmektedir.

Bağımsız değiken sayısı birden fazla ise, her bağımsız değişken ile kalıntı arasında ayrı ayrı sıra korelasyon katsayısı hesaplanarak *t*-test uygulanır. Kısaca test her bağımsız değişken için tekrarlanır.

Örnek

1982-1995 yılları arasındaki Gelir (Y) ve Tüketim (C) verilerinden tahmin edilen regresyon fonksiyonu aşağıda raporlanmıştır. Heteroskedasitenin varlığını, Spearman Sıra korelasyon testi ile araştıralım ($\alpha = 0,05$).

$$C_i = 1796.09 + 0.69Y_i \qquad n = 14 \quad R^2 = 0.92 \quad \hat{\sigma} = 328.78$$

$$Se(\hat{\beta}_i)\,(596.84)(0.056)$$

Modelin verileri ve kalıntılar aşağıdaki gibidir.

Yıllar	Tüketim(C)	Gelir (Y)	$\hat{u}_i$
1982	6702	7357	-177.21
1983	6812	7752	-340.13
1984	7399	8325	-149.02
1985	8095	8987	89.56
1986	8935	9699	437.64
1987	9447	10076	689.16
1988	9369	10365	411.48
1989	9097	10323	168.5
1990	8788	10213	-64.49
1991	8843	10637	-302.44
1992	9189	11126	-294.3
1993	9585	11485	-146.34
1994	9957	12051	-165.41
1995	10409	12693	-156.98

Spearman sıra korelasyon testi uygulamasının ilk aşaması kalıntılar ve bağımsız değişkenlere küçükten büyüğe veya büyükten küçüğe sıra numarası vermektir. Biz uygulamamızda küçükten büyüğe sıra numarası vereceğiz. Okuyucu isterse büyükten küçüğe sıra numarası vererek aynı sonuca ulaşabilir.

Gelir (Y)	$\hat{u}_i$	d_i	d_i^2
1	8	-7	49
2	11	-9	81
3	4	-1	1
4	2	2	4
5	13	-8	64
6	14	-8	64
9	12	-3	9
8	7	1	1
7	1	6	36
10	10	0	0
11	9	2	4
12	3	9	81
13	6	7	49
14	5	9	81
		$\sum d_i = 0$	$\sum d_i^2 = 524$

Not: $\sum d_i = 0$ olmalıdır.

Temel ve alternatif hipotezler

$$H_0 : \rho_s = 0$$

$$H_1 : \rho_s \neq 0$$

Spearman sıra korelasyon katsayısı

$$r_s = 1 - 6\left[\frac{\sum d_i^2}{n\left(n^2 - 1\right)}\right] = 1 - 6\left[\frac{524}{14\left(14^2 - 1\right)}\right] = -0.1516$$

olarak hesaplanır. Spearman sıra korelasyon katsayısı test istatistiğinin değeri:

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} = \frac{(-0,1516)\sqrt{14-2}}{\sqrt{1-(-0,1516)^2}} = -0,5312$$

$t = -0,5312$ değeri $t_c = t_{(0.05;14-2)} = t_{(0.05;12)} = \pm 2.179$ ile karşılaştırıldığında

$t_h = -0.5312 > t_c = -2.179$ veya $|t_h = -0.5312| < t_c = 2.179$ olduğu için tüketim modeli için homoskedasite varsayımının geçerli olduğu sonucuna varılır. Bağımsız değişken gelir ile kalıntıların mutlak değerleri arasında sistematik bir ilişki yoktur.

Bartlett Testi

Heteroskedasitenin araştırıldığı modele ilişkin veri seti birbirinden bağımsız k sayıda alt örneğe bölünür. Bu alt örneklerin normal dağılan anakütleden çekildiği varsayılmaktadır. Her alt ana kütle için ana kütle hata terimi varyansı tahmin edilir. Bunların serbestlik dereceleri $v_1, v_2, \ldots v_k$ 'a eşitttir. Burada dikkat edilmesi gereken husus hata terimi varyansının tahmininde en çok benzerlik tahmincisinin($\hat{\sigma}^2 = \sum \hat{u}^2 / n$) kullanılıyor olmasıdır.

Amaç bu alt örneklerin aynı anı kütleden gelip gelmediklerinin test edilmesidir. Bunun için temel ve alternative hipotezler aşağıdaki gibi kurulur.

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \cdots = \sigma_k^2 = \sigma^2$ Homoskedasite

$H_1 : \sigma_1^2 \neq \sigma_2^2 \neq \sigma_3^2 \neq \cdots \neq \sigma_k^2$ Heteroskedasite

Her alt örnek hata teriminin varyansı aynı anakütle hata teriminin bir tahminidir. Eğer temel hipotez doğru ise, aşağıdaki formül ana kütle hata terimi varyansının (σ^2) ortak bir tahmininin verecektir.

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k} v_i \hat{\sigma}_i^2}{\sum v_i}$$

Burada v_i alt örnek hacmidir. Temel hipotez A/B oranı ile test edilir. A/B oranı k-1 serbestlik derecesi ile χ^2 dağılımı gösterir. Burada A ve B aşağıdaki gibi hesaplanmaktadır.

$$A = v \log \hat{\sigma}^2 - \sum (v_i \log \hat{\sigma}_i^2)$$

$$B = 1 + \frac{1}{3}(k-1)\left[\sum\left(\frac{1}{v_i}\right) - \left(\frac{1}{v}\right)\right]$$

Test istatistiği:

$$\chi^2 = \frac{A}{B}$$

Test istatistiğinin hesaplanan değeri (χ^2_{hes}) tablo değerinden büyük ise temel hipotez reddedilir, alternative hipotez kabul edilir. Alt örneklerin hata terimi varyans tahminleri istatistiksel açıdan eşit değildir ve modelde heteroskedasite vardır.

Örnek

Aşağıdaki verilerden 1991-1997 ve 1998-2002 olmak üzere iki alt örnek elde ederek Barlett testi ile homoskedasite varsayımını araştırın.

Yıllar	**Y**	**X**	$\hat{u}_i$
1991	297	331	-3,34
1992	304	333	1,94
1993	308	338	2,12
1994	325	360	0,46
1995	340	378	0,08
1996	338	376	0,98
1997	358,5	398	1,62
1998	358	410	-8,64
1999	378	418	5,75
2000	392	446	-4,98
2001	413	462	2,29
2002	433	487	1,71

k_1		
Yıllar	$\hat{u}_i$	$\hat{u}_i^2$
1991	-3,34	11,1556
1992	1,94	3,7636
1993	2,12	4,4944
1994	0,46	0,2116
1995	0,08	0,0064
1996	0,98	0,9604
1997	1,62	2,6244
	$\sum \hat{u}_i^2 = 23,2164$	
$\hat{\sigma}_1^2 = 23,2164/7 = 3,32$		

k_2		
Yıllar	$\hat{u}_i$	$\hat{u}_i^2$
1998	-8,64	74,6496
1999	5,75	33,0625
2000	-4,98	24,8004
2001	2,29	5,2441
2002	1,71	2,9241
	$\sum \hat{u}_i^2 = 141,6807$	
$\hat{\sigma}_1^2 = 141,6807/5 = 28,34$		

Temel ve alternatif hipotezler

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$

$H_1 : \sigma_1^2 \neq \sigma_2^2$

k= 2 alt örnek $\quad v = \sum v_i = 7 + 5 = 12$

Ana kütle varyansının tahmini

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k} v_i \hat{\sigma}_i^2}{\sum v_i} = \frac{(7 \times 3,32) + (5 \times 28,24)}{12} = 13,75$$

$$A = v \log \hat{\sigma}^2 - \sum (v_i \log \hat{\sigma}_i^2)$$

$$A = (12 \times \log 13,75) - \left[(7 \times \log 3,32) + (5 \times \log 28,24) \right] = 2,74962$$

$$B = 1 + \frac{1}{3}(k-1) \left[\sum \left(\frac{1}{v_i} \right) - \left(\frac{1}{v} \right) \right]$$

$$B = 1 + \frac{1}{3}(2-1) \left[\left(\frac{1}{7} + \frac{1}{5} \right) - \left(\frac{1}{12} \right) \right] = 1,0865$$

Test istatistiği:

$$\chi^2 = \frac{A}{B} = \frac{2,74962}{1,0865} = 2,53$$

$\alpha = 0,05$ anlamlılık düzeyinde $k - 1 = 2 - 1 = 1$ serbestlik derecesiyle tablo değeri $\chi_1^2 = 3,84$.

$\chi_{hes}^2 = 2,53 < \chi_1^2 = 3,84$ olduğu için temel hipotez kabul edilir. Hata terimin varyansı gözlemler itibariyle sabittir, model homoskedastiktir.

The Goldfeld-Quandt Testi

Goldfeld-Quandt testi büyük örneklere uygulanan *F*-testidir. Bu test hata terimi varyansının bağımsız değişkenlerden biri ile pozitif ilişkili olduğunu varsayar. Genellikle

$\sigma_i^2 = \sigma^2 X^2$

Oolduğunu varsayar. σ_i^2, X ile pozitif (aynı yönde) ilişkilidir ve σ_i^2 X'in karesi ile orantılıdır. X'in değeri artıkça σ_i^2'ın değerleri de artmaktadır. Test rassal hataların normal dağıldığı ve birbirinden bağımsız oldukları, diğer bir ifade ile otokorelasyonun bulunmadığı varsayımına dayanmaktadır.

Goldfeld-Quandt testi, olası farklı varyanslı iki grup veri için tasarlanmıştır. Bu durumu tanıtmak için, ücretliler için tüketim fonksiyonu ile serbest meslek sahiplerinin tüketim modelinin farklılaştığı iktisadi gerçeklere uygundur. Serbest meslek sahipleri genel olarak ücretlilere göre daha fazla gelir elde ettikleri için tüketimleri ücretlilere göre farklılık arz eder. Ücretliler ve serbest meslek sahiplerini tek bir grup olarak düşünür ve ortalama tüketim denklemini elde etmek istersek, tüketim kalıplarının farklılığından dolayı değişen varyans problemi ile karşılaşmamız kaçınılmazdır.

Sorduğumuz soru: ücretlilerin tüketim varyansı ile karşılaştırıldığında serbest meslek sahiplerinin tüketim varyansı ne kadardır? Varyanslar aynı ya da farklı mıdır? Ücretlilerin tüketim ve serbest meslek sahiplerinin tüketim varyansı tüm gözlemler için sabit değildir. Goldfeld-Quandt testi değişen varyansın örneklemin iki gruba bölünebildiği –bu durumda ücretliler ve serbest meslek- biçimini test etmek amacıyla tasarlanmıştır ve varyansın iki grup için farklı olabileceğinden şüpheleniriz.

Test, her bir gruptan tahmin edilen hata varyanslarının karşılaştırılmasına dayanmaktadır. Goldfeld-Quandt testi, rassal hataların normal dağıldığı ve birbirleri ile ilişkisiz oldukları (otokorelasyon olmadığı) varsayımına dayanmaktadır.

Testin uygulamasının ilk aşamasında X bağımsız değişkeni küçükten büyüğe ilgili Y bağımlı değişkeni ile sıralanır. Sıralanan gözlemlerin ortasından c sayıda gözlem çıkarılır. n >30 ise c gözlem sayısının ¼'ü, n<30 ise c gözlem sayısının 1/5'dir. Geriye kalan n-c sayıdaki gözlem ikiye ayrılır $(n-c)/2$. Bu durumda ilk grupta küçük değerli X'ler, ikinci grupta ise daha büyük değerli X'ler yer alacaktır. Amaç bu iki grubun varyanslarının eşit olup olamadığını tespit etmektir. Eğer eşit değilse X'lerin değeri artarken rassal hatanın da varyansının arttığı sonucunu çıkarırız. Burada önemli iki önemli husus vardır. Bunlardan ilki dışlanan gözlem sayısı c'nin kaç olacağıdır. Monte Carlo denemelerine göre n=30 iken c=4, n=60 iken c=10 alınmasının yeterli olduğu konusunda bir görüş vardır. İkincisi ise çok değişkenli regresyon modelinde sıralama işleminin hangi X değişkenine göre yapılacağıdır. Şayet hangi bağımsız değişkenin değişen varsa sebep olabileceği öngörüle biliniyorsa sıralama o değişkene göre yapılır, aksi takdirde her değişken için test tekrarlanır.

İki alt gruptan küçük X değerlerinin bulunduğu gözlemler için rassal hatanın varyansı $\sigma_K^2 = Var(u_{iK}^2)$ ile, büyük X değerlerinin bulunduğu gözlemler için rassal hatanın varyansı $\sigma_B^2 = Var(u_{iB}^2)$ ile gösterilsin. Temel ve alternatif hipotez aşağıdaki gibidir.

$$H_0 : \sigma_K^2 = \sigma_B^2$$

$$H_0 : \sigma_K^2 \neq \sigma_B^2$$

$\sigma_K^2 = Var(u_{iK}^2)$ ve $\sigma_B^2 = Var(u_{iB}^2)$ eşitliklerinin söz konusu olduğu şeklindeki $\sigma_K^2 = \sigma_B^2$ sıfır hipotezini test etmek istiyoruz. Test istatistiği aşağıdaki gibidir.

$$F = \frac{\sum \hat{u}_B^2 / (n_B - k)}{\sum \hat{u}_K^2 / (n_K - k)} \sim F_{(n_B - k, n_k - k)}$$

Burada $(n_B - k)$ ve $(n_K - k)$, iki alt örneklemli regresyonlar için serbestlik dereceleridir ve $n_B = n_K$ eşitliği geçerlidir. Diğer bir deyişle bir varyans tahmininin onun gerçek ana kitle değeri oranına eşit bir paya ve diğer varyans tahmininin onun ana kitle değeri oranına eşit bir paydaya sahip olan F istatistiği, $(n_B - k, n_K - k)$ şeklindeki serbestlik derecesi ile *F* dağılımına sahiptir.

H0 doğru ise yukarıdaki test istatistiği aşağıdaki gibi olur.

$$F = \frac{\hat{\sigma}_B^2}{\hat{\sigma}_K^2}$$

Denklemde yer alan rassal hataların varyanslarının $\hat{\sigma}^2 = \sum \hat{u}_i^2 / n - k$ ile tahmin edildiğini hatırlayın. Hesaplanan test istatistiği verilen bir anlamlılık düzeyinde $(n_B - k, n_K - k)$ şeklindeki serbestlik dereceli F_c değerinden büyükse, heteroskedasite olduğu sonucuna varılır.

Bağımsız değişken sayısı birden fazla ise gözlemlerin büyüğe doğru sıralanması herhangi bir *X* değişkenine göre yapılabilir. Şayet hangi *X* göre sıralama yapılacağı konusunda tereddüt varsa her *X* değişkeni için test tekrarlanır.

Örnek:

30 aile için aşağıdaki örnek regresyon fonksiyonu elde edilmiştir.

$$Y_i = 9.29 + 0.64X \qquad n = 30 \quad R^2 = 0.95 \quad \hat{\sigma} = 9.18$$

$$Se(\hat{\beta}_i)(5.231)(0.638)$$

Yukarıdaki tüketim modelinde heteroskedasitenin varlığını Goldfeld-Quandt testi araştıralım.

Gözlem ler	**Y (Tüketim)**	**X (Gelir)**	**Sıralı Y**	**Sıralı X**
1	55	80	55	80
2	65	100	70	85
3	70	85	75	90
4	80	110	65	100
5	79	120	74	105
6	84	115	80	110
7	98	130	84	115
8	95	140	79	120
9	90	125	90	125
10	75	90	98	130
11	74	105	95	140
12	110	160	108	145
13	113	150	113	150
14	125	165	110	160
15	108	145	125	165
16	115	180	115	180
17	140	225	130	185
18	120	200	135	190
19	145	240	120	200
20	130	185	140	205
21	152	220	144	210
22	144	210	152	220
23	175	245	140	225
24	180	260	137	230
25	135	190	145	240
26	140	205	175	245
27	178	265	189	250
28	191	270	180	260

29	137	230	178	265
30	189	250	191	270

Yukarıdaki tablonun 4. ve 5. Sütunları Goldfeld-Quandt testinin birinci aşamasıdır. *X* değerleri küçükten büyüğe doğru ilgili *Y* değerleri ile birlikte sıralanmıştır. *n*=30 olduğuna göre ortadan *c*=4 gözlem analiz dışı bırakılacaktır. Bunlar 14.,15., 16. ve 17. gözlemlere ilişkin verilerdir. Böylece verimiz aşağıdaki gibi iki alt örnekleme ayrılmıştır.

1.Alt Örneklem (Küçük X'ler için)			**2.Alt Örneklem (Büyük X'ler için)**		
Gözlemler	**Sıralı Y**	**Sıralı X**	**Gözlemler**	**Sıralı Y**	**Sıralı X**
1	55	80	18	135	190
2	70	85	19	120	200
3	75	90	20	140	205
4	65	100	21	144	210
5	74	105	22	152	220
6	80	110	23	140	225
7	84	115	24	137	230
8	79	120	25	145	240
9	90	125	26	175	245
10	98	130	27	189	250
11	95	140	28	180	260
12	108	145	29	178	265
13	113	150	30	191	270

Yukarıdaki iki alt örnek verilerinden örnek regresyon fonksiyonu tahmin edilir. Tahmin sonuçları aşağıda raporlanmıştır.

1.Alt Örnek

$$Y_i = 3.409 + 0.67X_i \qquad n_1 = 13 \quad R_1^2 = 0.89 \quad \hat{\sigma}_1 = 5.86 \quad \sum \hat{u}_{1i}^2 = 377.17$$

$$Se(\hat{\beta}_i)(0.074)(8.70)$$

2. Alt Örnek

$$Y_i = -28.03 + 0.79X \qquad n_2 = 13 \quad R_2^2 = 0.77 \quad \hat{\sigma}_2 = 11.82 \quad \sum \hat{u}_{2i}^2 = 1536.8$$

$$Se(\hat{\beta}_i)(30.62)(0.132)$$

Goldfeld-Quandt testi için temel va alternatif hipotezler aşağıda verilmiştir.

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

1.Alt Örnek için rassal hataların varyansının tahmini aşağıdaki gibidir. Gözlem sayısı n_1 =13, tahmin edilen parametre sayısı k=2

$$\hat{\sigma}_1^2 = \frac{\sum \hat{u}_{1i}^2}{n_1 - k} = \frac{377.17}{13-2} = \frac{377.17}{11} = 34.29$$

2.Alt Örnek için rassal hataların varyansının tahmini aşağıdaki gibidir. (Gözlem sayısı n_2=13, Tahmin edilen parametre sayısı k=2

$$\hat{\sigma}_2^2 = \frac{\sum \hat{u}_{2i}^2}{n_2 - k} = \frac{1536.8}{13-2} = \frac{1536.8}{11} = 139.71$$

değerlerini elde ederek aşağıdaki F istatistiğini hesaplarız.

$$F = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2} = \frac{139.71}{34.29} = 4.07$$

Varyansların artabileceğini ancak gelirle azalmayacağını düşünerek %5 anlamlılık düzeyinde kritik değeri $F_{(11,11)} = 2.82$ olan bir tek kuyruk testi kullanırız. 4.07>2.82 olduğu için sabit varyansı ifade eden sıfır hipotezi varyansların gelirle birlikte artacağını ifade eden alternatif hipotez lehine reddedilir.

Park testi

Park testine göre her gözlemin hata terimin varyansı (σ_i^2) bağımsız değişken X'in bir fonksiyonudur. Bu ilişkinin fonksiyonel kalıbı ise aşağıdaki gibidir.

$$\sigma_i^2 = \sigma^2 X_i^{\beta} e^{v_i}$$

$$\ln \sigma_i^2 = \ln \sigma^2 + \beta \ln X_i + v_i$$

Burada $\ln \sigma_i^2$ bilinmediğiiçin yerine $\hat{u}_i^2$ geçer ve $\ln \sigma^2$ yerine α yazarak model aşağıdaki gibi yeniden yazılır.

$$\ln \hat{u}_i^2 = \alpha + \beta \ln X_i + v_i$$

Model EKK yöntemi ile tahmin edilir. Modelin hata terimi v_i için EKK yönteminin temel varsayımları yerine gelmeyebilir. Temel ve alternatif hipotezler

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

β istatistiksel açıdan anlamlı ise yani alternatif hipotez kabul edilirse, heteroskedasite söz konusudur.

Örnek

$\hat{Y}_i = 7,34 + 57,52 X_i$ *Y*=İthalat *X*= GSMH *n*=32

Yukarıda tahmin edilen modelde heteroskedasitenin varlığını araştırmak amacıyla aşağıdaki tardımcı regresyon tahmin edilmiştir. Heteroskedasitenin varlığı/yokluğu hakkında ne söyleyebilirsiniz.

$$\ln \hat{u}_i^2 = 3,69 + 0,38 \ln X_i$$

t (2,5574)(0,7066)

$\ln \hat{u}_i^2 = \alpha + \beta \ln X_i + v_i$ ise;

$H_0 : \beta = 0$ Homoskedasite

$H_1 : \beta \neq 0$ Heteroskedasite

$\alpha = 0,05$ *sd*=*n*-*k*=32-2=30 $t_c = t_{(0.05;30)} = \pm 2,042$ 'e eşittir. $t_h = 0,7066 < t_c = 2,042$ olduğu için temel hipotez reddedilemez. Hata teriminin varyansı *X* bağımsız değişkeninin aldığı değere göre değişiklik göstermemektedir. Homoskedasite varsayımı yerine gelmiştir.

Breusch-Pagan-Godfrey Testi

Goldfeld-Quandt testinin başarısı sadece ortadan atılan değişken sayısı *c*'nin değerine değil, aynı zamanda gözlemleri sıralamada doğru *X* değişkeninin tanımlamaya da bağlıdır. Breusch-Pagan-Godfrey testi bu mahsurları ortadan kaldırmak için önerilmiştir.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{k-1} X_{ik-1} + u_i$$

$$\sigma_i^2 = f(Z_{i1}, Z_{i2}, Z_{i3}, \ldots, Z_{i(s-1)})$$

Burada σ_i^2, stokastık(rassal olmayan) *Z*'lerin doğrusal fonksiyonudur. *Z*'ler, *X*'lerin tümü veya bazılarıdır.

$$\sigma_i^2 = \alpha_0 + \alpha_1 Z_{i1} + \alpha_2 Z_{i2} + \ldots + \alpha_{(s-1)} Z_{i(s-1)}$$

$\alpha_1 = \alpha_2 = \ldots = \alpha_{(s-1)} = 0$ ise $\sigma_i^2 = \alpha_0$ olacak ve homoskedasite varsayımı gerçekleşmiş olacaktır. Buna göre temel ve alternatif hipotezler aşağıdaki gibidir.

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_{(s-1)} = 0$$

$$H_1 : \alpha_1 \neq \alpha_2 \neq \ldots \neq \alpha_{(s-1)} \neq 0$$

Testin aşamaları aşağıdaki gibidir.

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{k-1} X_{ik-1} + u_i$ EKKY ile tahmin edilir ve modelin kalıntıları ($\hat{u}_i$) hesaplanır. En çok benzerlik tahmincisi ile ana kütle hata teriminin varyansı tahmin edilir.

$$\tilde{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n}$$

Sonrasında aşağıdaki gibi p_i değerleri oluşturulur.

$$p_i = \frac{\hat{u}_i^2}{\tilde{\sigma}^2}$$

p_i değerlerinin bapımlı değişken olduğu aşağıdaki yardumcı regresyon modeli modeli yazılır.

$$p_i = \alpha_0 + \alpha_1 Z_{i1} + \alpha_2 Z_{i2} + \ldots + \alpha_{(s-1)} Z_{i(s-1)} + v_i$$

model tahmin edilerek, modelin regresyon ile açıklanabilen değişmesi ($\sum \hat{y}_i^2$) hesaplanır. Sonrasında aşağıda verilen test istatistiği hesaplanır.

$$\theta = \frac{1}{2} \sum \hat{y}_i^2$$

θ, yardımcı regresyondaki bağımsız değşken sayısı (*s*-1) serbestlik derecesiyle χ^2_{s-1} dağılımına uygundur. Verilen bir anlamlılık düzeyi için $\theta > \chi^2_{s-1}$ temel hipotez reddedilir, alternative hipotez Kabul edilir. Modelde heteroskedasite vardır.

Örnek

$\hat{Y}_i = 9,2903 + 0,6378X_i$ $\sum \hat{u}_i^2 = 2361,153$ $R^2 = 0,9466$ n=30

Breusch-Pagan-Godfrey testini kullanarak modelde heteroskedasitenn varlığını araştırın.

$$\tilde{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n} = \frac{2361,153}{30} = 78,7051$$

$p_i = \hat{u}_i^2 / \tilde{\sigma}^2$ eşitliğinden p_i ler hesaplanır ve aşağıdaki yardımcı regresyon tahmin edilir.

$$\hat{p}_i = -0,7426 + 0,0101Z_i \qquad \sum \hat{y}_i^2 = 10,428 \qquad R^2 = 0,18$$

$SE(\hat{\alpha}_i)$ (0,7529) (0,0041)

Temel ve alternatif hipotezler:

$$H_0 : \alpha_1 = 0$$

$$H_1 : \alpha_1 \neq 0$$

Test istatistiği:

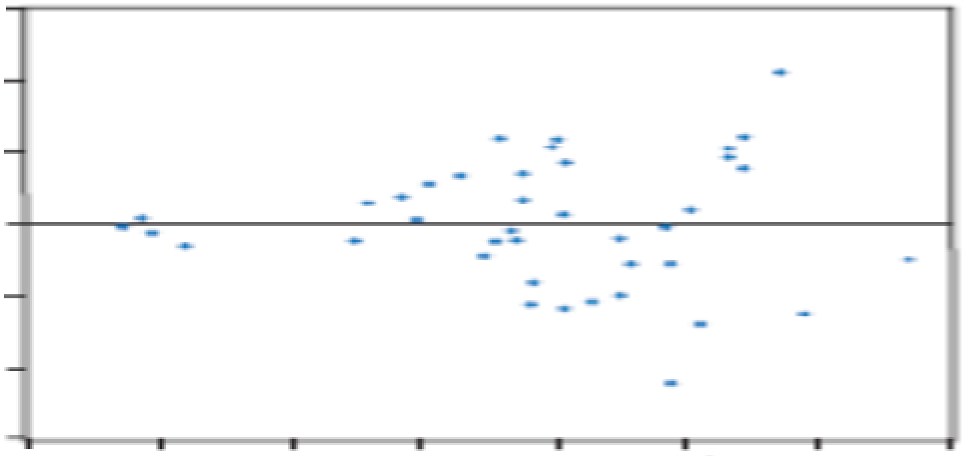
$$\theta = \frac{1}{2}\sum \hat{y}_i^2 = \frac{1}{2} \times 10,428 = 5,214$$

$\alpha = 0,05$ ve 1 serbestlik derecesi için tablo değeri $\chi_1^2 = 3,8414$ 'e eşittir. $\theta = 5,214 > \chi_1^2 = 3,8414$ olduğu için temel hipotez reddedilir, alternatif hipotez kabul edilir. Modelde heteroskedasite vardır.

Lagrange Çarpanı Testi için varyans fonksiyonu

Varyans fonksiyonu kavramına başlamak amacıyla öncelikle aşağıda verilen genel çoklu regresyon modeli için ortalama fonksiyonunu $E(Y)$ dikkate alalım.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{k-1} X_{ik-1} + u_i$$



En küçük kareler tüketim harcamaları kalıntılarının gelir düzeyine karşı grafiği.

Varyans fonksiyonu değişen varyansın varlığında geçerli bir yaklaşımdır. Varyansı $X_{i1}, X_{i2}, X_{i3}, \ldots, X_{i(k-1)}$dan farklı olarak $Z_{i1}, Z_{i2}, Z_{i3}, \ldots, Z_{i(k-1)}$ açıklayıcı değişkenler seti ile ilişkilendirmemizin dışında yukarıdaki çoklu regresyon modeli için ortalama fonksiyonunu benzerdir. Varyans fonksiyonunun genel bir biçimi aşağıdaki gibidir.

$$Var(Y_i) = \sigma_i^2 = E\left(u_i^2\right) = h\left(\alpha_0 + \alpha_1 Z_{i1} + \alpha_2 Z_{i2}, + \cdots + \alpha_k Z_{ik}\right)$$

h(.) fonksiyonuna ilişkin özel bir tanımlama yapmadığımız için bu genel bir biçimdir. Varyans fonksiyonuna göre Y_i 'nin varyansı Z'lere bağlı olarak her bir gözlem için değişecektir. Örneğin tüketim harcamaları örneği için ortalama ve varyans fonksiyonları sırasıyla aşağıdaki gibi verilmişti.

$$E(Y) = \beta_0 + \beta_1 X$$

$$Var\left(Y_i\right) = Var\left(u_i\right) = h\left(X_i\right)$$

Denklemlerde sadece bir tane *X* olduğu için bir tane *Z* olacaktır. Her ikisi de aynı değişken hane halkı geliridir.

Lagrange çarpanı testinin üssel fonksiyon, doğrusal fonksiyon gibi çok sayıda $h(.)$ fonksiyon için geçerli olması, değişen varyansın tespiti için tercih nedenidir. Biz dersimizde sadece doğrusal fonksiyonu ele alacağız.

Doğrusal varyans fonksiyonu aşağıdaki gibidir.

$$h\left(\alpha_0 + \alpha_1 Z_{i1} + \alpha_2 Z_{i2}, + \cdots + \alpha_{k-1} Z_{ik-1}\right) = \alpha_0 + \alpha_1 Z_{i1} + \alpha_2 Z_{i2}, + \cdots + \alpha_{k-1} Z_{ik-1}$$

$\alpha_1 = \alpha_2 = \alpha_3 = \cdots = \alpha_{k-1}$ olduğunda $h(.)$ fonksiyonunun aşağıdaki şekilde daralır.

$$h\left(\alpha_0 + \alpha_1 Z_{i1} + \alpha_2 Z_{i2}, + \cdots + \alpha_{k-1} Z_{ik-1}\right) = h\left(\alpha_0\right)$$

$h(\alpha_0)$ terimi bir sabittir ve $h(\alpha_0) = \alpha_0$ 'dır. Buna göre varyans herhangi bir açıklayıcı değişkene bağlı değildir. Diğer ifadeyle $\alpha_1 = \alpha_2 = \alpha_3 = \cdots = \alpha_{k-1} = 0$ ise değişen varyans söz konusu değildir; varyans sabittir ve söz konusu durumu $\sigma^2 = h(\alpha_0)$ şeklinde yazabiliriz.

Lagrange Çarpanı testi

Varyans fonksiyonu tanıtıldıktan sonra şimdi değişen varyansın testi için Lagrange Çarpanı testinin uygulama aşamalarını verebiliriz.

Lagrange Çarpanı testi için temel ve alternatif hipotezler aşağıdaki gibidir.

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_{s-1} = 0$$

H_1 :Tüm α 'ler sıfırdan farlıdır.

Temel ve alternatif hipotezler bir testin ilk bileşenleridir. Hipotezlerde yer alan α 'lar değişen varyansın kaynağı olduğu düşünülen değişkendir. Sonraki bileşen, bir test istatistiğidir. Bir test istatistiğini elde edebilmek için doğrusal varyans fonksiyonunu $Var(Y_i) = Var(u_i) = h(X_i)$ da yerine koyarak aşağıdaki eşitliğe ulaşılır.

$$Var(Y_i) = \sigma_i^2 = E\left(u_i^2\right) = h\left(\alpha_0 + \alpha_1 Z_{i1} + \alpha_2 Z_{i2}, + \cdots + \alpha_{s-1} Z_{is-1}\right)$$

$v_i = u_i^2 - E(u_i^2)$, bir hata karesi (u_i^2) ve onun ortalaması ($E(u_i^2)$) arasındaki fark ile tanımlanırsa yukarıdaki denklem, aşağıdaki biçimde yazılabilir.

$$u_i^2 = E(u_i^2) + v_i = \alpha_0 + \alpha_1 Z_{i1} + \alpha_2 Z_{i2}, + \cdots + \alpha_{s-1} Z_{is-1} + v_i$$

Varyans fonksiyonuna v_i 'nin eklenmesi ortalama fonksiyonuna u_i 'nin ilave edilmesiyle benzer bir amaca hizmet etmektedir. Bilindiği üzere $E(Y)$ ortalama fonksiyonuna u_i 'nin ilave edilmesiyle aşağıdaki genel regresyon modeli elde edilir:

$$Y_i = E\left(Y_1\right) + u_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{k-1} X_{ik-1} + u_i$$

Ancak önemli bir farklılık vardır. Genel regresyon modelinde bağımlı değişken Y gözlemlenebilir. Ancak , eğer $u_i^2 = E(u_i^2) + v_i = \alpha_0 + \alpha_1 Z_{i1} + \alpha_2 Z_{i2}, + \cdots + \alpha_{s-1} Z_{is-1} + v_i$ tahmin etmek istersek gerçek regresyon hataları (u_i) gözlemlenemediği için bilinmediğinden dolayı “bağımlı değişken” durumundaki u_i^2’in de bilinmediği sonucuna rahatlıkla ulaşılabilir. Ancak u_i^2’yi $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{k-1} X_{ik-1} + u_i$ ’’dan elde edilen en küçük kareler kalıntılarının kareleri (u_i^2) ile yer değiştirerek bu problem çözülür. Böylece

$$u_i^2 = \alpha_0 + \alpha_1 Z_{i1} + \alpha_2 Z_{i2}, + \cdots + \alpha_{s-1} Z_{is-1} + v_i$$

Yerine yardımcı regresyon olarak adlandırılan işlevsel biçimi aşağıdaki gibi yazılır.

$$\hat{u}_i^2 = \alpha_0 + \alpha_1 Z_{i1} + \alpha_2 Z_{i2}, + \cdots + \alpha_{s-1} Z_{is-1} + v_i$$

Önemli not: u_i^2 yerine $\hat{u}_i^2$’yi kullanmak yeni rassal hata v_i’nin tanımını değiştirir. Bu farklılığı göstermek için yukarıdaki denklemde esasen v_i yerine v_i^* yazılması gerekir. Ancak gereksiz karışıklıklardan kaçınmak için aynı notasyon kullanılacaktır.

Değişen varyans için varyans fonksiyonu testi için, yardımcı regresyon daki en küçük kareler tahmininden edinilen değerler kullanılır. $Z_{i1}, Z_{i2}, Z_{i3}, \ldots, Z_{is-1}$ değişkenlerinin $\hat{u}_i^2$’deki değişimi açıklamaya yardımcı olup olmadığını belirlemeyle ilgileniyoruz. Modelin tahmini ile elde edilen uyum iyiliği istatistiği R^2, *Z*’ler tarafından açıklanan $\hat{u}_i^2$’lerin değişim oranını ölçtüğü için bir test istatistiğinin genel elemanıdır. H_0 doğru ise, R^2 ile gözlem sayısının çarpımı $s-1$ serbestlik derecesiyle Ki-kare (χ^2) dağılımına sahiptir ve test istatistiği aşağıdaki gibidir.

$$\chi^2 = n \times R^2 \sim \chi_{s-1}^2$$

χ^2 dağılımını kullanmak için bir takım kısıtlar söz konusudur. χ^2 dağılımı, çok farklı türdeki hipotezlerin test edilmesi amacıyla kullanılan bir dağılımdır. *F* rassal değişkeni gibi χ^2 rassal değişkeni de sadece pozitif değerler alır. Büyük bir R^2 değeri sıfır hipotezine karşı kanıt sağladığı için (varyanstaki değişimlerin Z değişkenleri tarafından açıklandığını varsayar) test istatistiği için ret bölgesi dağılımın sağ kuyruğundadır. Böylece %5 anlamlılık düzeyi için H_0 reddedilir ve genellikle $\chi^2 \geq \chi^2_{(0,095;s-1)}$ ise sabit varyans varsayımının geçerli olmadığı değişen varyansın söz konusu olduğuna karar verilir. Lagrange çarpanı testi büyük örneklem testidir.

White Testi

White testi normallik varsayımına dayanmaz ve büyük örneklem testidir. Lagrange çarpanı testiyle ilgili problem, eğer değişen varyansın alternatif hipotezi doğru ise varyans fonksiyonunda bulunan değişkenler hakkında bilgi sahibi olunduğunu, diğer bir ifade ile bu değişkenlerin ($Z_{i1}, Z_{i2}, Z_{i3}, \ldots, Z_{i(s-1)}$) heteroskedasitenin kaynağı olduğunun öngörülmesidir. Gerçekte ise değişen varyans ilgili değişkenlere ilişkin herhangi bir bilgi olmaksızın test edilmek istenir. Genellikle, varyansı etkileyen değişkenler ortalama fonksiyonundakilerle aynıdır. Ortalama fonksiyonunun iki açıklayıcı değişkene sahip olduğunu varsayalım.

$$E\left(Y_1\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$$

White testinde Z 'ler X 'lere, X 'lerin karelerine ve X 'lerin olası çapraz çarpım terimlerine eşit olarak tanımlanır. White testi çarpım terimleri olmaksızın aşağıdaki şekilde gösterilir:

$$Z_1 = X_1 \quad , \quad Z_2 = X_2 \quad , \quad Z_3 = X_1^2 \quad , \quad Z_4 = X_2^2 \quad , \quad Z_5 = X_1 X_2$$

White testi, F testi ya da $\chi^2 = n \times R^2$ testi kullanılarak uygulanır. İki tane bağımsız değişkeni olan bir model için White testinin aşamaları aşağıdaki gibidir.

Regresyon denklemi tahmin edildikten $(\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2})$ sonra kalıntılar ($\hat{u}_i$) hesaplanır.

Yardımcı regresyon denklemi kurulur. Yardımcı regresyonun bağımlı değişkeni (Lagrange çarpanı testinde olduğu gibi) ilk aşamada hesaplanan kalıntıların kareleri, bağımsız değişkenleri ise yukarıda tanımlanan Z 'dir. Buna göre yardımcı regresyon aşağıdaki gibi tanımlanır.

$$\hat{u}_i^2 = \alpha_0 + \alpha_1 Z_{i1} + \alpha_2 Z_{i2} + \alpha_3 Z_{i3} + \alpha_4 Z_{i4} + \alpha_5 Z_{i5} + v_i$$

Yardımcı regresyonun belirginlik katsayısı (R^2) hesaplanır ve temel ve alternatif hipotezler kurulur.

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$$

$$H_1 : \alpha_1 \neq \alpha_2 \neq \alpha_3 \neq \alpha_4 \neq \alpha_5 \neq 0$$

Temel ve alternatif hipotezde yer alan parametreler Z'lerin katsayısıdır, sabit parametre temel ve alternatif hipoteze dahil edilmez.

H_0 doğru ise , R^2 ile gözlem sayısının çarpımı yardımcı regresyondaki bağımsız değişken sayısı *k* (burada *k*=5) serbestlik derecesiyle Ki-kare (χ_k^2) dağılımına sahiptir ve test istatistiği aşağıdaki gibidir.

$$\chi^2 = n \times R^2 \sim \chi^2$$

değeri hesaplanır. Gözlem sayısı hem ana modelde hem de yardımcı regresyon modelinde aynıdır.

Hesaplanan test istatistiğinin değeri (χ^2) ve χ^2 ile mukayese edilir. $\chi^2 > \chi_k^2$ ise sabit varyans varsayımını ifade eden H_0 temel hipotezi reddedilir. Modelin hata teriminin varyansı gözlemler itibariyle değişmektedir.

Örnek

White testinin uygulaması için tüketim harcamaları verileri kullanılacaktır. Hatırlanacağı üzere Örnek 2 verilerinden örnek regresyon fonksiyonu $\hat{Y}_i = 9.37 + 0.65X_i$ olarak elde edilmişti. Varyansın potansiyel olarak gelirin bir fonksiyonu olduğu tüketim harcamasının bu örneğinde değişen varyansı test etmek için $\sigma_i^2 = h\left(\alpha_0 + \alpha_1 X_i\right)$ şeklindeki varyans fonksiyonunda $H_0 : \alpha_1 = 0$ hipotezini alternatifi olan $H_1 : \alpha_1 \neq 0$ hipotezine karşı test ederiz.

$H_0 : \alpha_1 = 0$

$H_1 : \alpha_1 \neq 0$

$\hat{Y}_i = 9.37 + 0.65X_i$ modeli için belirginlik katsayısı $R^2 = 0,90$, n=10 olduğuna göre test istatistiği

$$\chi^2 = n \times R^2 = 10 \times 0.90 = 9$$

Sıfır hipotezinde sadece bir parametre olduğu için χ^2 testinin serbestlik derecesi bir'dir. $\alpha = 0,05$ için kritik değeri ise 3.84'tür.Test istatistiğinin değeri 9, tablo değeri 3.84'ten daha büyük olduğu için H_0 hipotezini reddederiz ve varyansın gelire bağlı değiştiğini kabul ederiz.

White testinin uygulamasına yardımcı regresyon $\hat{u}_i^2 = \alpha_0 + \alpha_1 X_i + \alpha_2 X_i^2$ fonksiyonunu EKK yaklaşımı ile tahmin ederek başlarız ve $H_0 : \alpha_1 = \alpha_2 = 0$ hipotezini $H_1 : \alpha_1 \neq 0$ veya $\alpha_2 \neq 0$ hipotezine karşı test ederiz.

$H_0 : \alpha_1 = \alpha_2 = 0$

$H_1 : \alpha_1 \neq 0$ veya $\alpha_2 \neq 0$

Yardımcı regresyon aşağıdaki gibi tahmin edilmiştir.

$$\hat{u}_i^2 = 1038.16 - 15.95X_i + 0.057X_i^2 \qquad R^2 = 0.85$$

Buradan test istatistiği

$$\chi^2 = n \times R^2 = 10 \times 0.85 = 8.5$$

olarak hesaplanır. $\alpha = 0,05$ için kritik değer $\chi^2_{(0.95,2)} = 5.99$'dir. Tekrar, gelire bağlı değişen varyansın söz konusu olduğu sonucuna ulaştık.

Örnek

$$\ln Y_i = \beta_0 + \beta_1 \ln X_{i1} + \beta_2 \ln X_{i2} + u_i \qquad n = 41$$

Y = İth+İhr Vergi/Toplam Kamu Geliri

X_1 = İth+İhr/GSMH

X_2 = Kişi başına GSMH

Yardımcı regresyon

$$\hat{u}_i^2 = -584 + 2,56 \ln X_{i1} + 0,69 \ln X_{i2} - 0,4 \ln X_{i1}^2 - 0,5 \ln X_{i2}^2 + 0,06 \ln X_{i1} \ln X_{i2} \qquad R^2 = 0,1148$$

Temel ve alternatif hipotezler

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$$

$$H_1 : \alpha_1 \neq \alpha_2 \neq \alpha_3 \neq \alpha_4 \neq \alpha_5 \neq 0$$

Test istatistiği

$$\chi^2 = n \times R^2 = 41 \times 0.1148 = 4,7068$$

$\alpha = 0,05$ için 5 serbestlik dercesi ile kritik değer $\chi^2_{(5)} = 11,0705$'dir. Hesaplanan test istatistiğinin değeri $\chi^2 = 4,7068 < \chi^2_{(5)} = 11,0705$ olduğu için temel hipotez reddedilemez, model homoskedastiktir. Hata terimi için sabit varyans varsayımı yerine gelmiştir.

Glejser Testi

Glejser testi; büyük örneklere uygulanan bir testtir ve Park testine benzer. Heteroskedasitenin varlığının araştırıldığı modelin kalıntıları ($\hat{u}_i$) hesaplanır. Alternatif regresyon modellerinden, $\hat{u}_i$'ların mutlak değerlerinin $\hat{\sigma}_i^2$ ile ilişkide bulunduğu düşünülen X değişkenine göre regresyonunu bulmayı önerir. Alternatif regresyon modelleri

$$|\hat{u}_i| = \beta_0 + \beta_1 X_i + v_i$$

$$|\hat{u}_i| = \beta_0 + \beta_1 \sqrt{X_i} + v_i$$

$$|\hat{u}_i| = \beta_0 + \beta_1 \frac{1}{X_i} + v_i$$

$$|\hat{u}_i| = \beta_0 + \beta_1 \frac{1}{\sqrt{X_i}} + v_i$$

$$|\hat{u}_i| = \sqrt{\beta_0 + \beta_1 X_i} + v_i$$

$$|\hat{u}_i| = \sqrt{\beta_0 + \beta_1 X_i^2} + v_i$$

Yukarıdaki son iki denklem EKK yöntemi ile tahmin edilemez.

Örnek

$\hat{Y}_i = 7,34 + 57,52 X_i \qquad n = 32$

$Y =$ İhracat $\qquad X =$ GSMH

$|\hat{u}_i| = 4,94 + 14,07 X_i$
$t \qquad (4) \quad (2,61)$

$\alpha = 0,05$ için n-k=32-2=30 serbestlik derecesiyle tablo değeri $t_c = 2,042$ 'e eşittir.

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

$t_h = 2,61 > t_c = 2,042$ olduğu için, altternatif hipotez reddedilemez ve *X* heteroskedasitenin nedenidir.

$|\hat{u}_i| = 2,0073 + 15,22 \sqrt{X_i}$
$t \qquad (0,99) \quad (2,74)$

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

$t_h = 2,74 > t_c = 2,042$ olduğu için, altternatif hipotez reddedilemez ve *X* heteroskedasitenin nedenidir.

Heteroskedasitenin Düzeltilmesi: Heteroskedasitenin Varlığı Durumunda Parametre Tahmin Yöntemleri

Herhangi bir heteroskedasite testi sonucunda heteroskedasitenin saptanması halinde, bunun düzeltiebilmesi işlemi verilerin uygun bir dönüşüm ile homoskedastik hale getirilmesidir. Hetroskedasitenin varlığı halinde, gerekli düzeltme yapılarak, daha isabetli ve daha tesirli parametre tahmini yapabilmek için iki ana yöntem vardır. Bunlar:

1) Genelleştirilmiş En Küçük Kareler
2) Ağırlıklı (tartılı) En Küçük Kareler

Genelleştirilmiş En Küçük Kareler

Heteroskedasitenin varlığı halinde EKK yöntemiyle tahmin edilen parametreler eğilimsiz olmakla birlikte etkin değildirler. EKK yöntemi her gözleme eşit ağırlık vermektedir. Genelleştirilmiş EKK yöntemi ise bilgiyi dikkate alır ve BLUE tahminler üretebilir.

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Ana hütle hata terimnin gözlemler itibariyle değeri biliniyorsa (σ_i^2) model aşağıdaki gibi dönüştürülür.

$$\frac{Y_i}{\sigma_i} = \beta_0^* \frac{X_{0i}}{\sigma_i} + \beta_1^* \frac{X_i}{\sigma_i} + \frac{u_i}{\sigma_i}$$ burada X_{0i} her i değeri için $X_{0i} = 1$

Dönüştürülmüş model

$$Y_i^* = \beta_0^* X_{0i}^* + \beta_1^* X_i^* + u_i^*$$

Dönüştürülmüş modelin hata teriminin (u_i^*) varyansı ise

$$Var(u_i^*) = E(u_i^*)^2 = E\left(\frac{u_i}{\sigma_i}\right)^2$$

$$E(u_i^*)^2 = \frac{1}{\sigma_i^2} E\left(u_i^2\right)$$ burada $E(u_i^2) = \sigma_i^2$

$$Var(u_i^*) = E(u_i^*)^2 = 1$$

Dönüştürülmş modelin hata teriminin varyansı 1'e eşit olup sabittir. Dönüştürülmüş modelin ağırlıklandırılmış kalıntıkarelerinin toplamı ise

$$\sum \hat{u}_i^{*2} = \sum (\text{Y}_i^* - \beta_0^* X_{0i}^* - \beta_1^* X_i^*)^2$$

Veya

$$\sum\left(\frac{\hat{u}_i}{\sigma_i}\right)^2 = \sum\left[\left(\frac{Y_i}{\sigma_i}\right)-\beta_0^*\left(\frac{X_{0i}}{\sigma_i}\right)-\beta_1^*\left(\frac{X_i}{\sigma_i}\right)\right]^2 \qquad \text{burada } w_i = 1/\sigma_i \text{ olmak üzere}$$

Ağırlıklandırılmış kalıntıkarelerinin toplamı minumuma indirilerek, aşağıdaki gibi Genelleştirilmiş EKK tahmincisi β_1^* elde edilir.

$$\beta_1^* = \frac{(\sum w_i)\left(\sum w_i X_i Y_i\right)-\left(\sum w_i X_i\right)\left(\sum w_i Y_i\right)}{(\sum w_i)\left(\sum w_i X_i^2\right)-\left(\sum w_i X_i\right)^2}$$

Varyansı ise

$$Var(\beta_1^*) = \frac{\sum w_i}{(\sum w_i)\left(\sum w_i X_i^2\right)-\left(\sum w_i X_i\right)^2}$$

Ağırlıklı (tartılı) En Küçük Kareler

Düzeltme işleminde iki seçenekten hareket edilebilir. Bunlardan birincisi ana kütle hata terimi varyanslarının (σ_i^2) bilinmesi veya tahmin edilebilmesi hali, ikincisi ise bu durumun dışında olan $\sigma_i^2 = h(X)$ ilişkisine göre düzeltme işlemidir, diğer bir ifade ile σ_i^2 'nun bilimediği durumdur. Ağırlıklı EKK , genelleştirilmiş EKK yönteminin özel halidir.

σ_i^2 **bilindiği durum:**

Böyle bir durumun sadece kuramsal geçerliliği bulunmaktadır. Ekonometrik uygulamalarda σ_i^2 (bağımlı değişken ve hata teriminin varyansı) bilinmesi çok mümkün değildir, ancak tahmin edilebilir.

σ_i^2 biliniyor veya tahmin ediliyorsa, σ_i^2 'nın kare kökü standart sapma (σ_i) ile düzeltme işlemi yapılır.

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ $\qquad$ H_1 Kabul, modelde heteroskedasite var.

Bu durumda verilerin σ_i ile dönüştürüldüğü aşağıdaki rgresyon tahmin edilir.

$$\frac{Y_i}{\sigma_i} = \hat{\beta}_0^* \frac{1}{\sigma_i} + \hat{\beta}_1^* \frac{X_i}{\sigma_i} + \frac{\hat{u}_i}{\sigma_i}$$

Yukarıdaki modelde σ_i tartı olarak ele alınmaktadır. Yöntemde gözlemlerin (birimlerin) regresyon doğrusundan uzaklığı dikkate alınmaktadır. Gözlemler regresyon doğrusundan uzaklaştıkça σ_i^2 büyüyecek buna karşın $1/\sigma_i^2$ küçülecektir. Dolayısıyla regresyondan uzak olan gözlemlerin tartısı bağıl olarak küçük kalmaktadır. Böylece EKK yönteminin varyansı en büyük gözleme ağırlık vermesi sorunu ortadan kaldırılmaktadır.

Dönüştürlmüş modelde sabit terim yoktur, dolayısıyla β_0^* ve β_1^* parametrelerinin tahmini orjinden geçen regresyon modelinin tahminine dayanmaktadır. Ayrıca dönüştürlmüş modelde iki tane bağımsız değişken vardır, bunlar $1/\sigma_i$ ve X_i/σ_i 'dir.

σ_i^2 **bilinmediği durum:**

$Y_i = \beta_0 + \beta_1 X_i + u_i$

Heteroskedasitenin fonksiyonel biçimine göre;

1. Durum: Hata teriminin varyansı X_i^2 ile doğru orantılıdır.

$E(u_i^2) = \sigma^2 X_i^2$

Bu durumda model X_i ' ye bölünerek dönüştürülür.

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$\frac{Y_i}{X_i} = \beta_0 \frac{1}{X_i} + \beta_1 \frac{X_i}{X_i} + \frac{u_i}{X_i}$$

$$\frac{Y_i}{X_i} = \beta_0 \frac{1}{X_i} + \beta_1 + v_i$$

Burada v_i, dönüştürülmüş modelin hata terimi olup, varyansı aşağıdaki gibidir.

$$E(v_i^2) = E\left(\frac{u_i}{X_i}\right)^2 = \frac{1}{X_i^2} E(u_i^2) \qquad E(u_i^2) = \sigma^2 X_i^2 \quad \text{olduğuna göre}$$

$$E(v_i^2) = \frac{1}{X_i^2}(\sigma^2 X_i^2) = \sigma^2$$

$$E(v_i^2) = \sigma^2$$

Böylece dönüştürülmüş modelin hata teriminin (v_i) varyansı sabittir. Aşağıdaki model EKK yöntemi ile tahmin edilir.

$$\frac{Y_i}{X_i} = \beta_0 \frac{1}{X_i} + \beta_1 + v_i$$

$Y_i = \beta_0 + \beta_1 X_i + u_i$ modeline dönmek için tahmin edilen model X_i ile çarpılır.

2. Durum: Hata teriminin varyansı X_i ile doğru orantılıdır.

$$E(u_i^2) = \sigma^2 X_i$$

Bu durumda dönüştürme işlemi X_i'in kare kökü ile yapılır.

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$\frac{Y_i}{\sqrt{X_i}} = \beta_0 \frac{1}{\sqrt{X_i}} + \beta_1 \frac{X_i}{\sqrt{X_i}} + \frac{u_i}{\sqrt{X_i}}$$

$$\frac{Y_i}{\sqrt{X_i}} = \beta_0 \frac{1}{\sqrt{X_i}} + \beta_1 \sqrt{X_i} + v_i$$

Burada $X_i > 0$ olmalıdır. Dönüştürülmüş modelin hata teriminin varyansı

$$E(v_i^2) = E\left(\frac{u_i}{\sqrt{X_i}}\right)^2 = \frac{1}{X_i} E(u_i^2) \qquad E(u_i^2) = \sigma^2 X_i$$

$$E(v_i^2) = \frac{1}{X_i}(\sigma^2 X_i) = \sigma^2$$

$$E(v_i^2) = \sigma^2$$

Dönüştürülmüş model EKK yöntemi uygulanarak tahmin edilir. Burada dikkat edilmesi gereken husus modelin orjinden geçen bir regresyon modeli olmasıdır. Model tahmin edildikten sonra $\sqrt{X_i}$ ile çarpılarak ilk modele dönülür.

3. **Durum:** Hata teriminin varyansı, bağımlı değişkenin ortalamasının karesiyle orantılıdır.

$$E(u_i^2) = \sigma^2 \left[E(\mathrm{Y}_i)\right]^2 \qquad E(Y) = \beta_0 + \beta_1 X_i$$

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$\frac{Y_i}{E(Y)} = \beta_0 \frac{1}{E(Y)} + \beta_1 \frac{X_i}{E(Y)} + \frac{u_i}{E(Y)}$$

$$\frac{Y_i}{E(Y)} = \beta_0 \frac{1}{E(Y)} + \beta_1 \frac{X_i}{E(Y)} + v_i$$

Burada $E(v_i^2) = \sigma^2$'dir. $E(Y)$, bilinmeyen ana kütle parametreleri β_0 ve β_1'e bağlıdır. Dolayısıyla yukarıdaki dönüştürme işelemi uygulanabilir değildir. Ancak örnek verilerinen $E(Y)$'nin tahmini $\hat{Y}_i$'dir. Dolayısıyla dönüştürme işlemi $\hat{Y}_i$'e göre yapılır. Büyük örneklerde başarı artar.

$$\frac{Y_i}{\hat{Y}_i} = \beta_0 \frac{1}{\hat{Y}_i} + \beta_1 \frac{X_i}{\hat{Y}_i} + v_i$$

4. Durum: Logaritmik Dönüşüm

$Y_i = \beta_0 + \beta_1 X_i + u_i$ modelinde yer alan değişkenlerin doğal logaritmaları alınarak modele dahil edilmesi ve yeni modelin EKK yöntemi ile tahminidir. Bu dönüşme işleminin yapılabilmesi için $Y_i > 0$ ve $X_i > 0$ olmalıdır.

$$\ln Y_i = \beta_0 + \beta_1 \ln X_i + u_i$$

Yöntemin sakıncaları

Çok değişkenli regresyon modellerinde dönüştürme işleminin hangi X_i değişkenine göre yapılacağı sorun teşkil etmektedir.

$Y_i = \beta_0 + \beta_1 X_i + u_i$ modelinde Y_i ile X_i ilişkisizken, dönüştürülmüş modeldeki değişkenler arasında ilişki vardır. Bu durum Sahte regresyon olarak bilinmektedir.

t ve *F* testleri büyük örnekler için geçerlidir.