LOJİSTİK REGRESYON ANALİZİ
(Logistic Regression Analysis)

Lojistik regresyon analizi sınıflama ve atama işlemi yapmaya yardımcı olan bir regresyon yöntemidir. **Normal dağılım varsayımı, süreklilik varsayımı ön koşulu yoktur.**

Açıklayıcı değişkenlere göre cevap değişkeninin beklenen değerleri olasılık olarak elde edilen bir regresyon yöntemidir.

Ayırma (Diskriminant) analizi verilerin sınıflandırılması ve belirli olasılıklara göre belirli sınıflara atanmasını sağlayan bir yöntemdir. **Veri setindeki değişkenlerin sınıflamaya etkilerini Ayrıma analizi belirlemek mümkündür. Fakat Ayrıma Analizi çok değişkenli normal dağılım varsayımını ön koşul kabul etmektedir**.

Günlük hayatta gözlenen fenomenlerin bazıları var-yok, başarılı-başarısız gibi **ikili**, bazıları yok-orta-çok, hiç-az-çok gibi **sıralı** veya ikiden çok isimsel ölçekle ölçülmüş meslek (işçi-memur-emekli), tedavi (RT,KT,RT+KT) gibi isimsel ölçekli veriler olabilir.

LOJİSTİK REGRESYON ANALİZİ
(Logistic Regression Analysis)

Bağımlı değişken ikili, sıralı veya isimsel ölçekli sonuçların ortaya çıkmasında hangi faktör(ler) etkilidir? Bu soruya cevap vermek için Normal dağılım varsayımı aramayan Lojistik Regresyon Analizinden yararlanılır.

Lojistik Regresyon Analizi bağımlı değişkenin tahmini değerlerini olasılık olarak hesaplayarak, olasılık kurallarına uygun sınıflama yapma imkanı veren bir yöntemdir.

DEĞİŞKENLER	DOĞRUSAL REGRESYON ANALİZİ	LOJİSTİK REGRESYON ANALİZİ
BAĞIMLI	SÜREKLİ SAYISAL KESİKLİ SAYISAL	NİTELİK
BAĞIMSIZ	SÜREKLİ SAYISAL KESİKLİ SAYISAL	SÜREKLİ SAYISAL KESİKLİ SAYISAL NİTELİK (Her bağımsız değişken başka bir ölçüm biçimine de sahip olabilir)

Nitelik bağımlı değişken:

2 Kategorili olabilir (Binominal)	**: İyileşti-iyileşmedi, yaşıyor-öldü, etkili- etkisiz gibi.**
2+ Kategorili sırasız olabilir (Multinominal)	: Çalışıyor, çalışmıyor, emekli gibi
2+ Kategorili sıralı olabilir : (Ordinal)	Çok etkili-orta derecede etkili-etkisiz gibi

Her durumda lojistik regresyon analizi uygulanabilir.

Bağımlı Değişken Kategori Sayısı	Bağımsız Değişken Sayısı	Bağımsız Değişkenin Kategori Sayısı	Uygulanacak Yöntem
2	1	2	Binominal lojistik regresyon
2	1	2+	Binominal lojistik regresyon
2	2+	Çeşitli	Çok değişkenli lojistik regresyon
2+ sırasız	Tek/çok	Çeşitli	Multinominal lojistik regresyon
2+ sıralı	Tek/çok	Çeşitli	Ordinal lojistik regresyon

Lojistik regresyon Yöntemleri

Lojistik regresyonda 3 temel yöntem vardır. İkili (binary), sıralı (ordinal) ve nominal lojistik regresyon yöntemleridir.

1.İkili Lojistik Regresyon (Binary Logistic Regression):

İkili cevap içeren bağımlı değişkenlerle yapılan lojistik regresyon analizidir. Açıklayıcı değişkenler faktör yada ortak değişkelerdir (covariate). Faktör değişkenler isimsel ölçekli kategorik değişken, ortak değişkenler ise sürekli değişken olmalıdır.

.

Lojistik regresyonda odds oranı kullanılır. Odds oranı (OR), olma olasılığının olmama olasılığına oranı olarak tanımlanır.

Odds : Odds başarı ya da görülme olasılığının "P", başarısızlık ya da görülmeme olasılığına "1-P" oranıdır. Odds değeri $(0,+\infty)$ arasında değerler alır.

$$Odds=P/(1-P)$$

Odds Ratio (OR): İki odds'un birbirine oranıdır. İki değişken arasındaki ilişkinin özet bir ölçüsüdür. Lojistik regresyonda **$OR=exp(\beta)$** olarak hesaplanır.

Odds oranı (OR) 1'e yakın değişkenler Y'nin değişimine önemli etkide bulunan etkenler değildir. Bu değişkenlerin katsayıları önemli değil ise "değişken önemli risk faktörü değildir" biçiminde yorumlanır.

1'den büyük OR değerleri (katsayı önemli olmak koşuluyla) etkenin önemli bir risk faktörü olduğu yorumu yapılır. Yani risk artış göstermektedir. OR 1'den küçükse risk azalmaktadır denilir. Sıfıra yakın değerler ise katsayı önemli olmak koşulu ile etkenin önemli bir risk faktörü olduğunu fakat Y'nin düşük değerler almasına neden olduğu negatif etkili bir faktör olduğunu belirtir.

OR'nin - Exp(b)- Yorumu

Değ	b	Sh(b)	Wald	P	Exp(b)
Cinsiyet (B/E)	-0,874	0,416	4,422	0,035	0,420
PEF	-0,015	0,006	6,578	0,010	0,985
Alerjik Hast. (Yok/Var)	0,891	0,489	3,314	0,046	2,436

Bağımlı değişken astım hastalarının iyileşip-iyileşmemesidir. (hastaların remisyonda olup-olmaması).

Cinsiyet **(nitel)** için Odds oranı 0,42 bulunmuştur. Referans grup Bayan alındığında erkeklerin hastalığın devam etmesi (iyileşmemesi) riski %58 (1-0,42) daha azdır. Referans grup olarak erkekler alınırsa, bayanların iyileşmeme riski (hastalığın devam etmesi) erkeklere göre 1/0,42=2,38 kat daha fazladır.

Modelde PEF değişkeni **nicel** bir değişkendir. PEF değişkeninin değeri 1 birim artığında hastalığın iyileşmeme riski %1,5 (1-0,985) azalmaktadır.

Alerjik **(nitel)** olanların hastaların iyileşmeme riski alerjik olmayanlara göre 2.4 kat daha fazladır.

Risk	Hastalık		Toplam
	Var	Yok	
Var	35	16	51
Yok	25	61	86
Toplam	60	77	137

Riskli olanlarda hastalığa yakalanma odds'u:

35/16= 2.18,

Risksiz olanlarda hastalığa yakalanma odds'u:

25/61= 0.41'dir.

Bu iki odds'un birbirine oranı odds ratio'yu verir:

Odds ratio=2.18/ 0.41 = 5.3 OR=a*d/b*c=5,3

Risk altında olanların hastalığa yakalanma riski, risk altında olmayanlara göre 5.3 kat daha fazladır.

Lojit Fonksiyon:

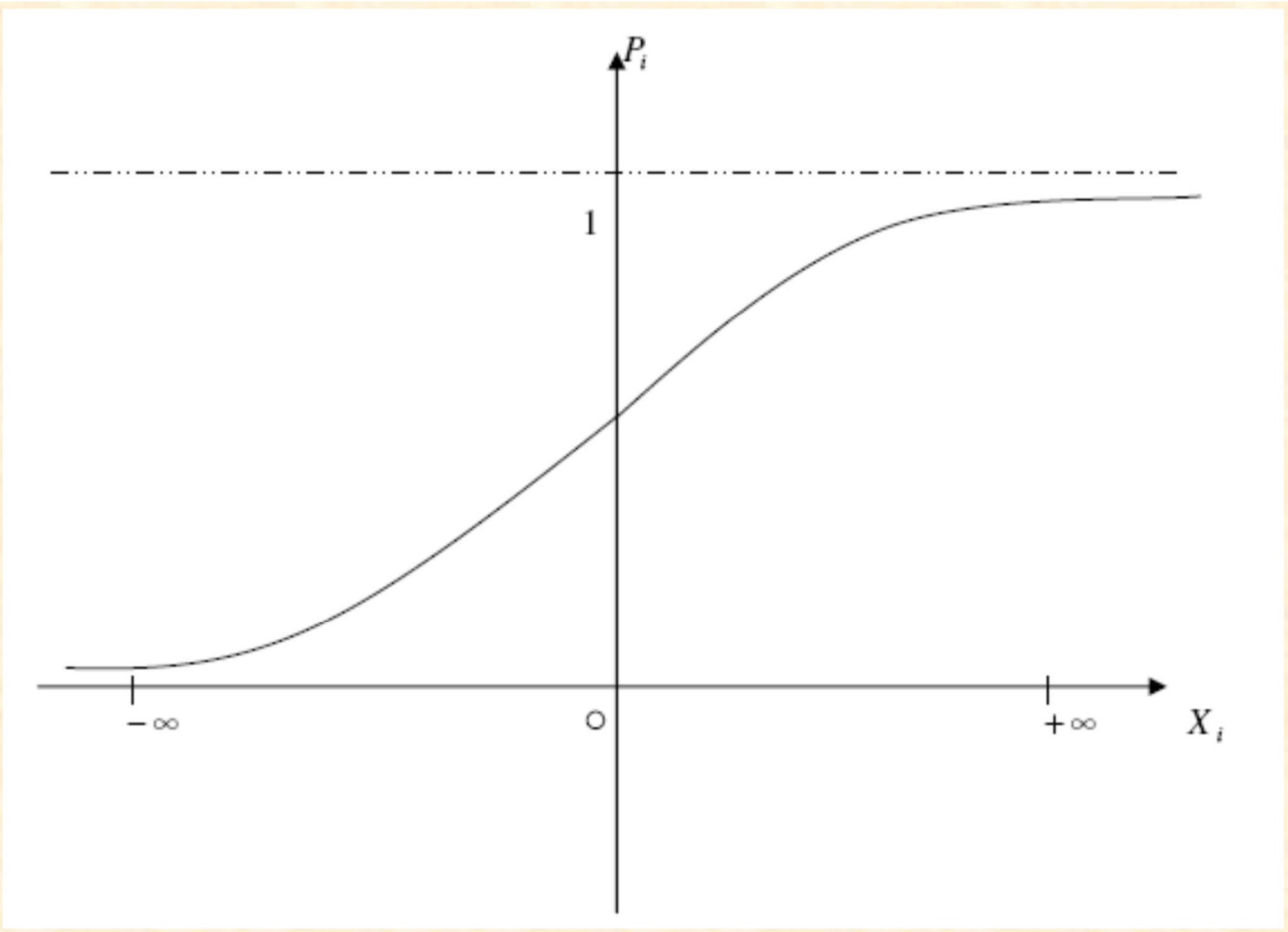
Odds ratio'nun doğal logaritmasıdır. Odds ratio asimetriktir. Doğal logaritması alınarak simetrik hale dönüştürülür.

$$OR = \frac{P}{1-P} \qquad Lojit(P) = \ln\left(\frac{P}{1-P}\right) = \ln(OR)$$

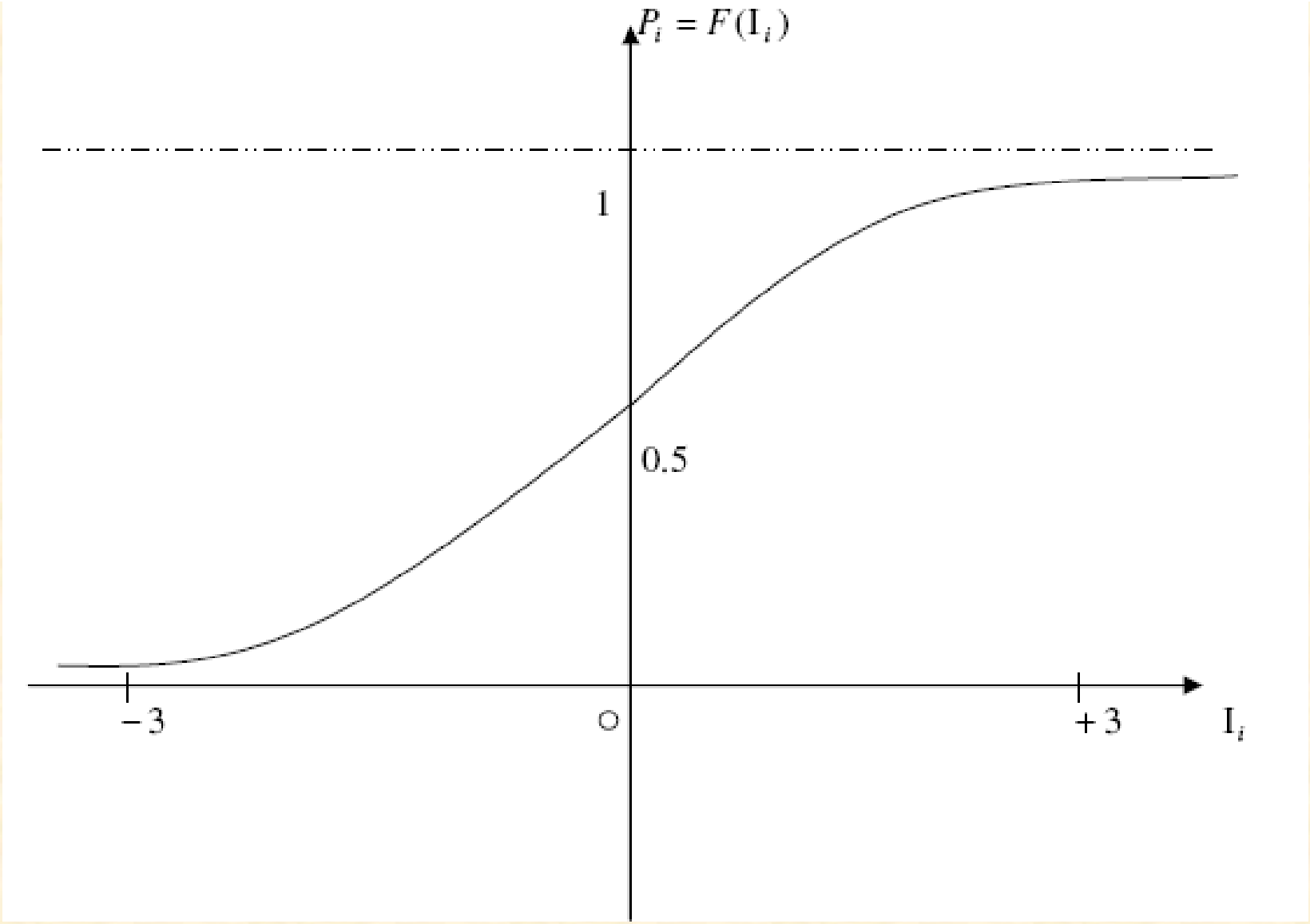
Bir olasılığın Odds değeri 0 ile +∞ arasında değerler alırken, aynı olasılığın lojit değeri -∞ ile + ∞ arasında değerler alır.

- P arttıkça lojit(P)'de artar.
- P<0.5 ise lojit(P) negatif,
- P>0.5 ise lojit(P) pozitif değerler alır.
- P, 0 ile 1 arasında iken lojit(p) reel sayılar doğrusu üzerinde değerler alabilir.

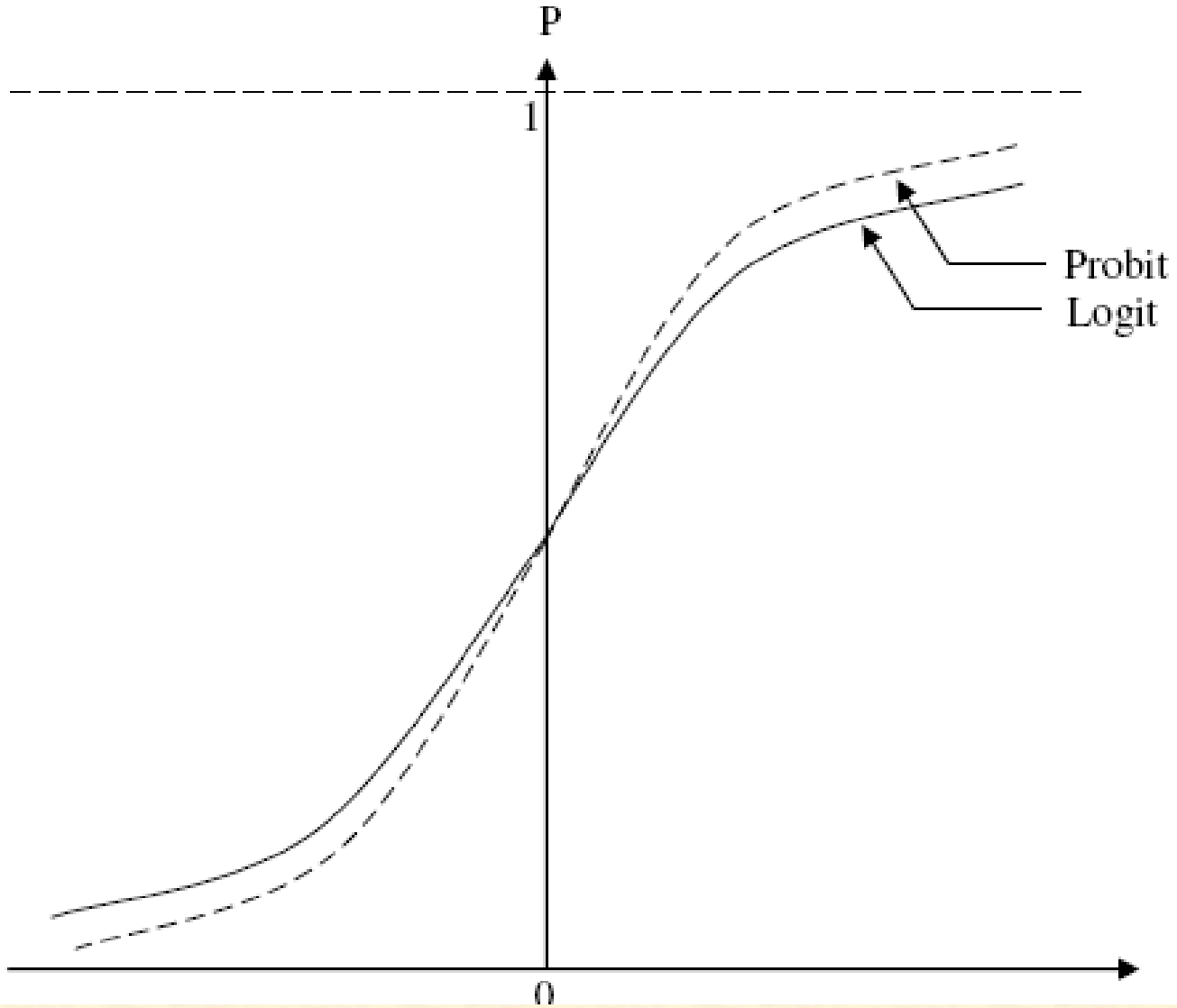
(Hosmer ve Lemeshow, 1989:307).



Logit Birikimli Dağılım Fonksiyonu



Probit Model



Logit ve Probit Model Birikimli Dağılım Eğrisi

Lojistik Regresyon Modeli:

$$z = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$P = \frac{e^z}{1 + e^z}$$

$$P = \frac{1}{1 + e^{-z}}$$

$$Lojit(P) = \ln\left(\frac{P}{1-P}\right) = \ln(e^z) = z = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

$$\frac{P}{1-P} = e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k} = OR$$

P: İncelenen olayın gözlenme olasılığı

Bağımlı değişken (Y)	Bağımsız değişken (X)	
	x=1	x=0
y=1	$P(1) = e^{\beta o + \beta 1} / 1 + e^{\beta o + \beta 1}$	$P(0) = e^{\beta o} / 1 + e^{\beta o}$
y=0	$1- P(1) = 1 / 1+ e^{\beta o + \beta 1}$	$1- P(0) = 1 / 1+ e^{\beta o}$
Toplam	1	1

Lojistik regresyon analizi sonucunda elde edilen modelin uygun olup olmadığı "model ki-kare" testi ile, her bir bağımsız değişkenin modelde varlığının anlamlı olup olmadığı ise Wald istatistiği ile test edilir.

Wald ki-kare istatistiği t değerlerinin karesine eşittir.

$$t = \frac{\hat{\beta}}{SH_{\hat{\beta}}} \quad , \quad t^2 = Wald$$

VARSAYIMLAR

Lojistik regresyon yönteminde doğrusal regresyon analizindeki gibi katı varsayımlar yoktur. Bu nedenle araştırıcılara önemli esneklik sağlamaktadır. Ancak aşağıdaki varsayımların dikkate alınması gerekir:

- Uygun tüm bağımsız değişkenler modele dahil edilmelidir
- Uygun olmayan tüm bağımsız değişkenler dışlanmalıdır
- Aynı birey üzerinde bir kez gözlem yapılmalı, tekrarlayan ölçümler olmamalıdır.
- Ölçüm hataları küçük olmalı, kayıp (eksik) veri olmamalıdır. Hatalar, katsayıların tahmininde yanlılığa ve modelin yetersizliğine neden olur.
- Bağımsız değişkenler arasında çoklu bağlantı (Multicollinearity) olmamalıdır.
- Aşırı değerler olmamalıdır.
- Örneklem büyüklüğü yeterli olmalıdır.
- Bağımlı değişkenin beklenen varyansı ile gözlenen varyansı arasında büyük bir fark varsa modelin yetersiz olduğu ve yeniden tanımlanması gerekir.

Sonuç değişkenin ikili değerler alması nedeni ile hata terimi sıfır ortalamalı ve P(1-P) varyanslı Binom dağılımına sahiptir.

X_i verildiğinde Y_i'nin koşullu beklenen değeri $E(Y_i|X_i)=\alpha+\beta X_i$

Y_i=1 olma olasılığı P_i
Y_i=0 olma olasılığı ise $(1-P_i)$

$$E(Y_i)=\sum Y_i P(Y_i)=0\times(1-P_i)+1\times(P_i)=P_i$$

$$e_i = Y_i - \alpha - \beta X_i$$

$$Y_i = 1 \quad \text{için} \quad e_i = 1 - \alpha - \beta X_i$$

$$Y_i = 0 \quad \text{için} \quad e_i = -\alpha - \beta X_i$$

Y_i	e_i	Olasılık
1	$1-\alpha-\beta Xi$	P_i
0	$-\alpha-\beta Xi$	$1-P_i$
Toplam		1

$$P_i = \begin{cases} \alpha+\beta X_i & 0 < \alpha+\beta X_i < 1 \\ 1 & \alpha+\beta X_i \geq 1 \\ 0 & \alpha+\beta X_i \leq 1 \end{cases}$$

$$E(e_i) = (1-\alpha-\beta X_i) * P_i + (-\alpha-\beta X_i)(1-P_i) = 0$$

$$P_i = \alpha + \beta X_i$$

$$1 - P_i = 1 - \alpha - \beta X_i$$

$E(e_i) = 0$

$$Var(e_i) = E(e^2) = (-\alpha-\beta_i X_i)^2(1-P_i) + (1-\alpha-\beta X_i)^2(P_i)$$

$$= (-\alpha-\beta i X_i)^2(1-\alpha-\beta_i X_i) + (1-\alpha-\beta_i X_i)^2(\alpha+\beta_i X_i)$$

$$= (\alpha+\beta_i X_i)(1-\alpha-\beta_i X_i) = P_i(1-P_i)$$

Lojistik Regresyon Analizi Adımları

1. Modele girecek değişkenler belirlenir. Bu amaçla önsel bilgiden ya da istatistiksel tekniklerden yararlanılabilir.
2. Modelin parametreleri tahmin edilir. Ardından **modelin tümünün anlamlılığı olabilirlik oranı ile test edilir.** Model anlamlı değilse analize son verilir. Eğer model anlamlı bulunursa diğer aşamaya geçilir.
3. Tahmin edilen model parametrelerinin tek tek anlamlılığı incelenir. Bu amaçla olabilirlik oranı ya da Wald istatistiği kullanılabilir. Her katsayının anlamlılığı incelendikten sonra, teklik oranları incelenerek, açıklayıcı değişkenlerin bağımlı değişken üzerindeki etkileri yorumlanabilir.
4. Tahmin edilen model parametreleri kullanılarak, her bir gözlemin hangi gruptan geldiği tahmin edilir.
5. **Modelin uyum iyiliğini** incelemek amacıyla doğru sınıflandırma yüzdesi ve Hosmer-Lemeshow ölçütleri kullanılır. Modelin uyum iyiliği kabul edilebilir düzeyde ise 5. aşamadaki grup tahminleri kullanılabilir. Aksi halde 2. aşamaya geçilerek modele girecek değişkenler yeniden gözden geçirilir ve işlemler tekrar edilir.

Uyum İyiliği İstatistikleri

Kurulan modelin uyum iyiliği testi Hosmer-Lemeshow'un hem onlu risk grupları hem de sabit kesim noktası yöntemine göre hesaplanmaktadır. Uyum iyiliğine karar vermek için onlu risk grupları yöntemine göre hesaplanmak isteniyorsa, Hosmer-Lemeshow

$$\hat{C}_g^* = \sum_{k=0}^{1}\sum_{i=1}^{10}\frac{(o_{ki} - e_{ki})^2}{e_{ki}}$$

Hosmer-Lemeshow istatistiği, t-2 serbestlik dereceli ki-kare dağılımı göstermektedir. $\hat{C}_g^* < \chi^2_{\alpha,(t-2)}$ ise model uyumunun oldukça iyi olduğu sonucuna varılır.

Kestirilen modelin uyum iyiliği testi sabit kesim noktası yöntemiyle hesaplanmak istendiğinde ise, Hosmer-Lemeshow $\hat{H}_g^*$ istatistiği kullanılmaktadır.

$$\hat{H}_g^* = \sum_{k=0}^{1}\sum_{i=1}^{10}\frac{(o'_{ki} - e'_{ki})^2}{e'_{ki}}$$

Hosmer-Lemeshow istatistiği, t-2 serbestlik dereceli ki-kare dağılımı göstermektedir. $\hat{H}_g^* < \chi^2_{\alpha,(t-2)}$ ise model uyumunun oldukça iyi olduğu sonucuna varılacaktır.

Hosmer-Lemeshow istatistiği Model Ki-kare istatistiği olarak da bilinir. Bu istatistik lojistik regresyon modelini genel olarak test eder. Yokluk hipotezi aşağıdaki gibi kurulur.

H_0: Sabit terim dışındaki tüm katsayılar sıfırdır.

Hosmer-Lemeshow istatistiği olabilirlik oran testidir ve modelde bağımsız değişkenlerin olmadığı $-2\log L_0$ istatistiği ile modelde bağımsız değişkenlerin yer aldığı $-2\log L$ istatistiği arasındaki fark alınarak hesaplanır. Bu istatistik incelenen modelin parametre sayısı ile sabit terimli modelin parametreleri arasındaki fark bir serbestlik dereceli ki-kare dağılımına uyar. **Modelin anlamlı olması arzu edilir ($p<0,05$).**

-2logL İstatistiği

Bu istatistik sapan ki-kare istatistiği olarak bilinir. **Bu istatistiğin anlamlı olmaması Lojistik regresyonda istenen durumu gösterir. -2logL istatistiği analize bağımsız değişken ilave edildiğinde modelin hatasını gösterir.** Bu nedenle -2logL istatistiği bağımlı değişkendeki açıklanmayan varyansın anlamlılığını gösterir.

Log olabilirlik (log likelihood) değeri 0-1 aralığında değerler almaktadır. Bu oran bağımlı değişkenin bağımsız değişkenler tarafından tahmin edilme olasılığını göstermektedir. 1'den küçük sayıların logaritması (0,-∞) arasındadır. logL istatistiği maksimum olabilirlik algoritması ile tahmin edilmektedir.

-2logL istatistiği yaklaşık olarak ki-kare dağılımına uyar. Lojistik regresyon analizinde -2logL istatistiği regresyon analizindeki hata kareler toplamına (HKT) benzer.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	31,989[a]	,438	,586
2	24,434[b]	,535	,716

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

b. Estimation terminated at iteration number 7 because parameter estimates changed by less than ,001.

Yukarıdaki tabloda her adımda modelin verileri nasıl temsil ettiği gösteren tahmini olasılıklarla gerçek olasılıklar arasındaki ilişki ve -2logL istatistikleri verilmiştir. Lojistik regresyonda bir sonraki adımda (step) bağımlı değişken ile bağımsız değişkenler arasındaki ilişkinin derecesini gösteren **Cox-Snell R^2 ile Nagelkerle R^2 değerlerinin yüksek olmasını** ve **-2logL istatistiğinin ise düşük olması istenir**. Modelin verileri tam olarak iyi temsil etmesi için olabilirlik 1 ve -2logL istatistiğinin ise sıfır olması gerekir.

Daha küçük -2logL değerine sahip model her zaman tercih edilmelidir.

Sözde R^2:

$$R^2 = 1 - (\ln L / \ln L_0)$$

L_0, sadece sabit terimin (β_0) yer aldığı modelin en çok olabilirlik değeridir. L ise tahmin edilen tüm parametrelerin yer aldığı modelin en çok olabilirlik değeridir.

R^2, değeri verideki belirsizliğin model tarafından açıklanabilen oranını göstermektedir. L=1 olduğunda, lnL=0 ve R^2=1 olur. Bu da ele alınan bağımsız değişkenler tarafından bağımlı değişkendeki değişimin tamamının açıklandığının ve modelin mükemmel olduğunun bir göstergesidir.

Cox-Snell R^2 ve Nagelkerke R^2

R^2 bağımlı değişkenin açıklanan varyansının yüzdesini gösterir ancak Lojistik regresyonda bağımlı değişkenin varyansı bu değişkenin olasılık dağılımına (frekans dağılımı) bağlıdır. İki gruplu bir bağımlı değişkenin varyansı grup frekansları eşit olduğu zaman (0,5*0,5=0,25) maksimum olur. Bu açıdan regresyon analizindeki R^2 ile lojistik regresyondaki R^2 farklıdır. **Lojistik regresyonda bağımlı ve bağımsız değişkenler arasındaki ilişkinin gücünün ölçülmesinde kullanılan iki sözde R^2 istatistiği vardır. Bu istatistikler, Cox-Snell R^2 istatistiği ve Nagelkerke R^2 istatistiğidir** (Kalaycı, 2005). Bu değerlerin >0.2 olması istenir.

$$R^2_{CS} = 1 - \left(\ln L_0 / \ln L\right)^{\frac{2}{n}}$$

$$R^2_N = \frac{R^2_{CS}}{1 - L_0^{\frac{2}{n}}}$$

Modelin Uygunluğunun Test Edilmesi

Oluşturulan birçok modelin uygunluğunun test edilmesi, değerlendirilmesi ve bu modeller arasından en uygun modelin seçilebilmesi için tüm gözlem değerlerini temsil edecek bir istatistiksel değere ihtiyaç duyulur.

Modelin uygunluğunun testi için Pearson ki-kare istatistiği, sapma ölçüsü ve sözde R^2 değeri yaygın olarak kullanılmaktadır. Ancak sözde R^2 değeri, kategorik bağımlı değişkenin yer aldığı modeller için kesin sonuçlar vermez (Long, 1997). Ayrıca gözlem sayısının az olduğu çalışmalarda sapma ölçüsü yetersiz kalmaktadır.

Model Ki-kare anlamlı ($p<0,05$), Cox-Snell R2 ve Nagelkerke R2 0,2'den büyükse modelin anlamlı olduğu söylenebilir. Hosmer-Lemeshow testinde ise $p>0,05$ ise modelin veriye uyumunun iyi olduğuna karar verilir.

Ki-kare ist.

$$\chi^2 = \sum\sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Sapma İst.

$$D = 2\sum\sum O_{ij} \ln\left(\frac{O_{ij}}{E_{ij}}\right)$$

O_{ij}: Gözlemlenmiş değerler
E_{ij}: Beklenen değerleridir

P>0,05 ise modelin uygun olduğu belirtilir.

Lojistik Regresyon Analizinde Model Uygunluğunun Değerlendirilmesi

Lojistik regresyon analizinde model uygunluğunun değerlendirilmesinde gerçek olasılıklar ile tahmin edilen olasılıklar arasındaki farklara bakılır.

1. Standart Olmayan Hatalar:

Standart olmayan hatalar (e_i) gerçek olasılıklar ile tahmin edilen olasılıklar arasındaki farktır. Logit hatalar aşağıdaki gibi bulunur.

$$Logit\ Hata_i = \frac{e_i}{P_i(1-P_i)}$$

2. Standart Hatalar

Standart hatalar , standart olmayan hataların (e_i) kendi standart sapmalarına bölünmesiyle elde edilir. Büyük örnekler için standart hatalar 0 ortalama 1 standart sapma ile normal dağılıma uyar.

$$Z_i = \frac{e_i}{\sqrt{P_i(1-P_i)}}$$

3. Sapma (Deviance) Değerleri

Her bir birimin sapma değerleri aşağıdaki gibi hesaplanır. **Büyük sapma değerleri modelin ilgili veriyi iyi temsil etmediğini gösterir**. Büyük örnekler için sapma değerleri normal dağılıma uyar.

Y_i=1 için $$Sapma(Deviance) = \sqrt{-2\ln(P_i)}$$

Y_i=0 için $$Sapma(Deviance) = -\sqrt{-2\ln(1-P_i)}$$

Sapma değerlerinin Q-Q ve trendsiz Normal P-P grafikleri çizilerek normallik varsayımı kontrol edilir.

4. Uzaklık (Leverage) Değerleri

Uzaklık değerleri tahmin edilen değerler üzerinde büyük etkisi olan birimlerin belirlenmesi amacıyla kullanılır. Uzaklık değerleri 0 (tamamen etkisiz) ve 1 (tamamen etkili) aralığı içinde değerler alır.

Uzaklık değerlerinin ortalaması p/n ile bulunur p sabit dahil modeldeki parametre sayısını, n ise gözlem sayısını gösterir. Böylece uzaklık değerleri ortalama uzaklık değeri ile karşılaştırılır.

5. Cook Uzaklığı (Cook's Distance)

Cook uzaklığı değeri herhangi bir birimin model üzerindeki etkisini gösterir. Cook uzaklığı belirli bir birimin modelden çıkartılması durumunda lojistik regresyon katsayılarının ne kadar değişeceğini gösterir.

$$Cook_i = Z_i^2 \left(\frac{h_i}{1 - h_i} \right)$$

Z_i: Standartlaştırılmış hatalar
h_i: Leverage uzaklık değerleri

6. DfBeta Değerleri

DfBeta değerleri belirli bir birimin modelden çıkartılması durumunda lojistik regresyon katsayılarının ne kadar değişeceğini gösterir. Sabit terim dahil her bir değişken için DfBeta değerleri aşağıdaki gibi hesaplanır.

$$DfBeta(\beta_0^i) = \beta_0 - \beta_0^i \qquad DfBeta(\beta_1^i) = \beta_1 - \beta_1^i$$

β_0 ve β_1 bütün birimler modele dahil edilmesi durumundaki parametreleri gösterir.

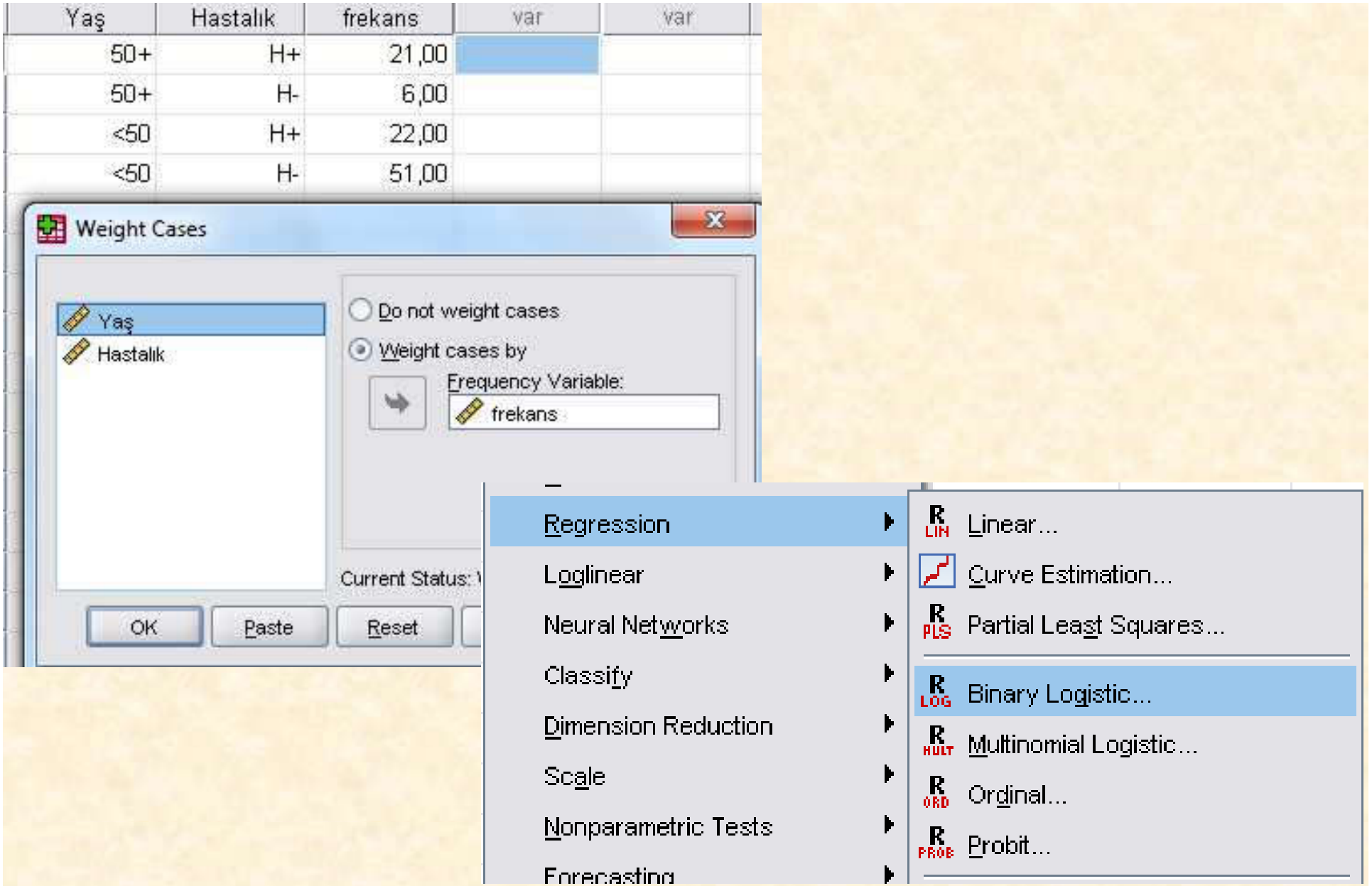
β_0^i ve β_1^i ise i.birimin modelden modelden çıkartılmasıyla hesaplanan parametreleri gösterir.

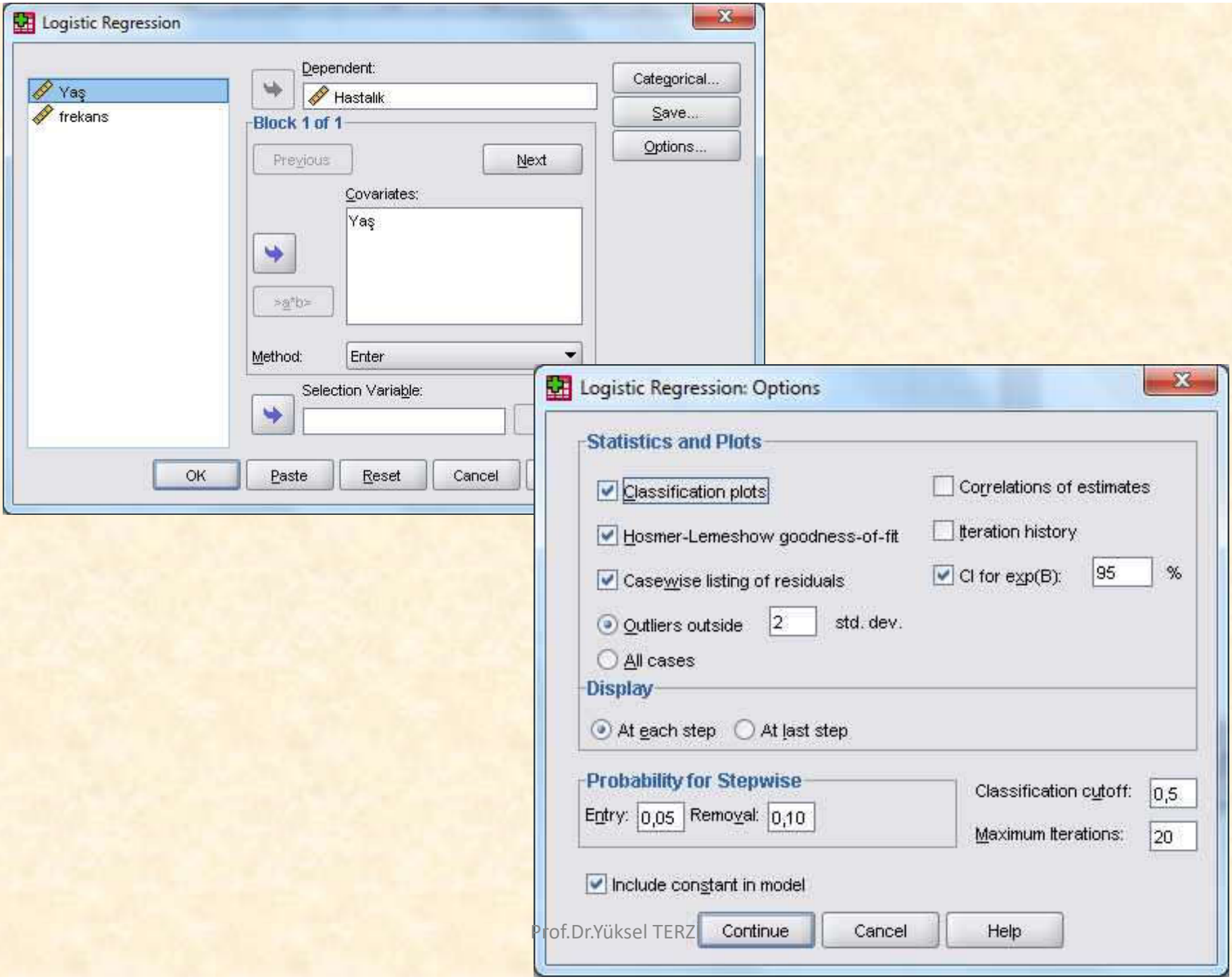
Her bir açıklayıcı değişken için bulunan DfBeta değerlerinin **Q-Q grafikleri çizilerek normallik varsayımı kontrol edilir.**

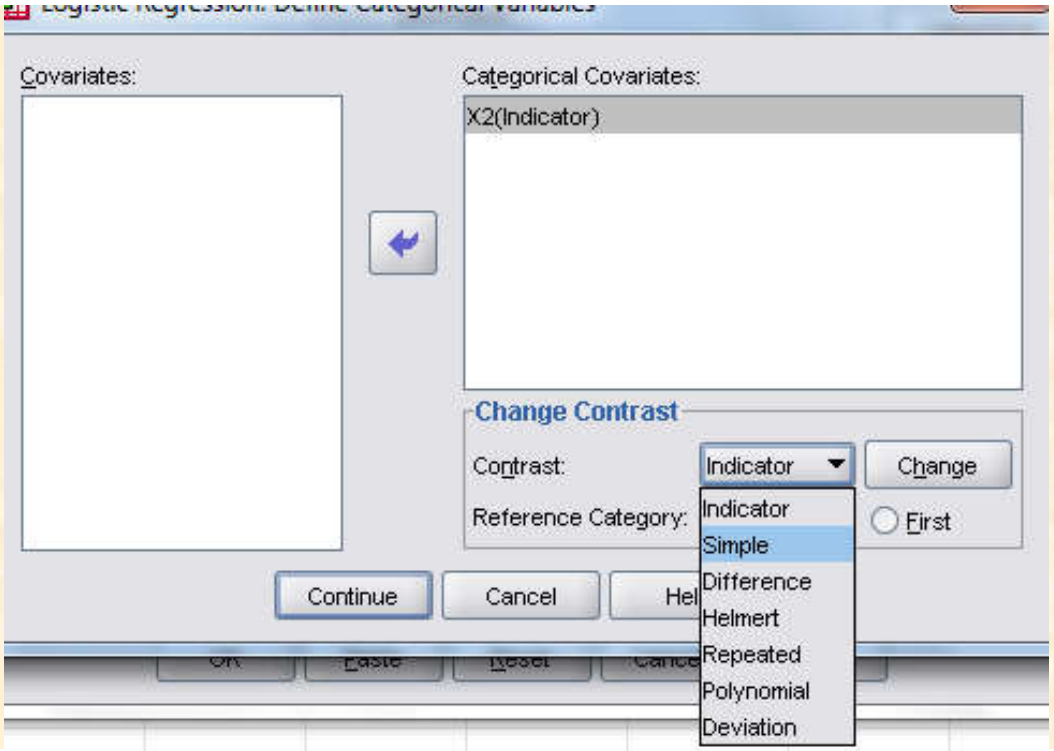
Örnek. “Hastalığa yakalanma” ile “Yaş” ile arasındaki ilişkiyi lojistik regresyonla inceleyelim:

Bağımlı değişken : Hastalığa yakalanma
Bağımsız değişken: Yaş

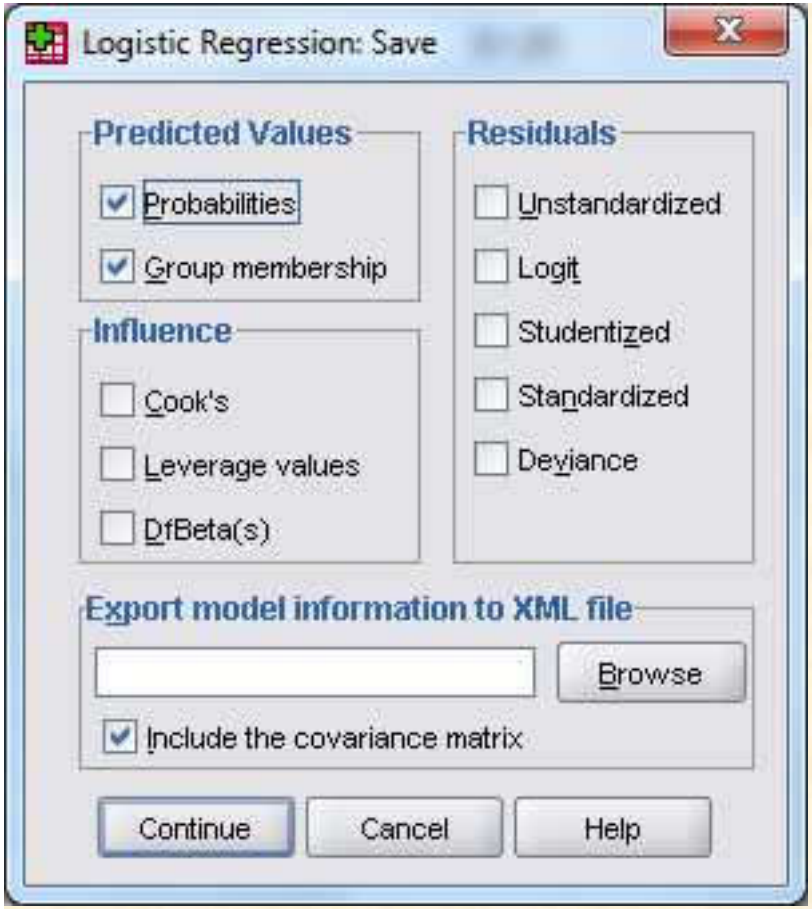
Risk (yaş)	Hastalık		Toplam
	Var	Yok	
50+	21	6	27
<50	22	51	73
Toplam	43	57	100


Yaş
Hastalık
frekans
var
var
50+ H+ 21,00
50+ H- 6,00
<50 H+ 22,00
<50 H- 51,00
Weight Cases
Yaş
Hastalık
Do not weight cases
Weight cases by
Frequency Variable:
frekans
Current Status:
OK
Paste
Reset
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Linear...
Curve Estimation...
Partial Least Squares...
Binary Logistic...
Multinomial Logistic...
Ordinal...
Probit...





1. ***Tekrarlı (repeated) Yöntem***: Herbir grubun riskini, kendinden öncekine göre hesaplanmasına olanak sağlar.
2. ***Fark (difference) Yöntemi***: Herbir grubun riskini, kendinden öncekilerin ortalama riskine göre hesaplanmasını sağlar.
3. ***Helmert Yöntemi***: Herbir grubun riskini, kendinden sonrakilerin ortalama riskine göre hesaplanmasını sağlar.
4. ***Sapma (deviance) Yöntemi***: Herbir grubun riskini, grupların toplam riskine göre hesaplanmasını sağlar.
5. ***Basit (simple) Yöntem***: Herbir grubun riskini, referans (temel) gruba göre hesaplanmasını sağlar.



Yaş	Hastalık	frekans	PRE_1	PGR_1
50+	H+	21,00	0,22222	H+
50+	H-	6,00	0,22222	H+
<50	H+	22,00	0,69863	H-
<50	H-	51,00	0,69863	H-

Classification Table[a]

Observed			Predicted		
			Hastalık		
			H+	H-	Percentage Correct
Step 1	Hastalık	H+	21	22	48,8
		H-	6	51	89,5
	Overall Percentage				72,0

a. The cut value is ,500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Yaş	2,0935	0,5285	15,6899	1	,000	8,114	2,880	22,861
	Constant	-3,3463	0,9603	12,1424	1	,000	,035		

a. Variable(s) entered on step 1: Yaş.

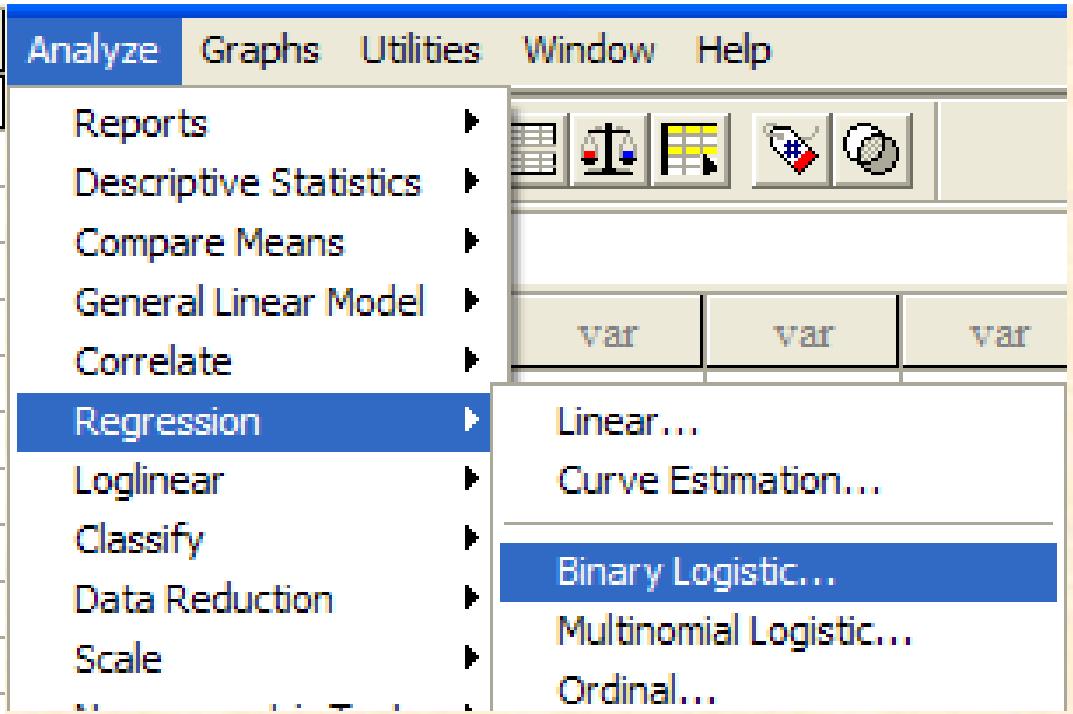
$$Z_{\beta_1} = \frac{\hat{\beta}_1}{SH_{\hat{\beta}_1}} = \frac{2,0935}{0,5285} = 3,961 \quad , \; Wald = Z^2 = 3,961^2 = 15,69$$

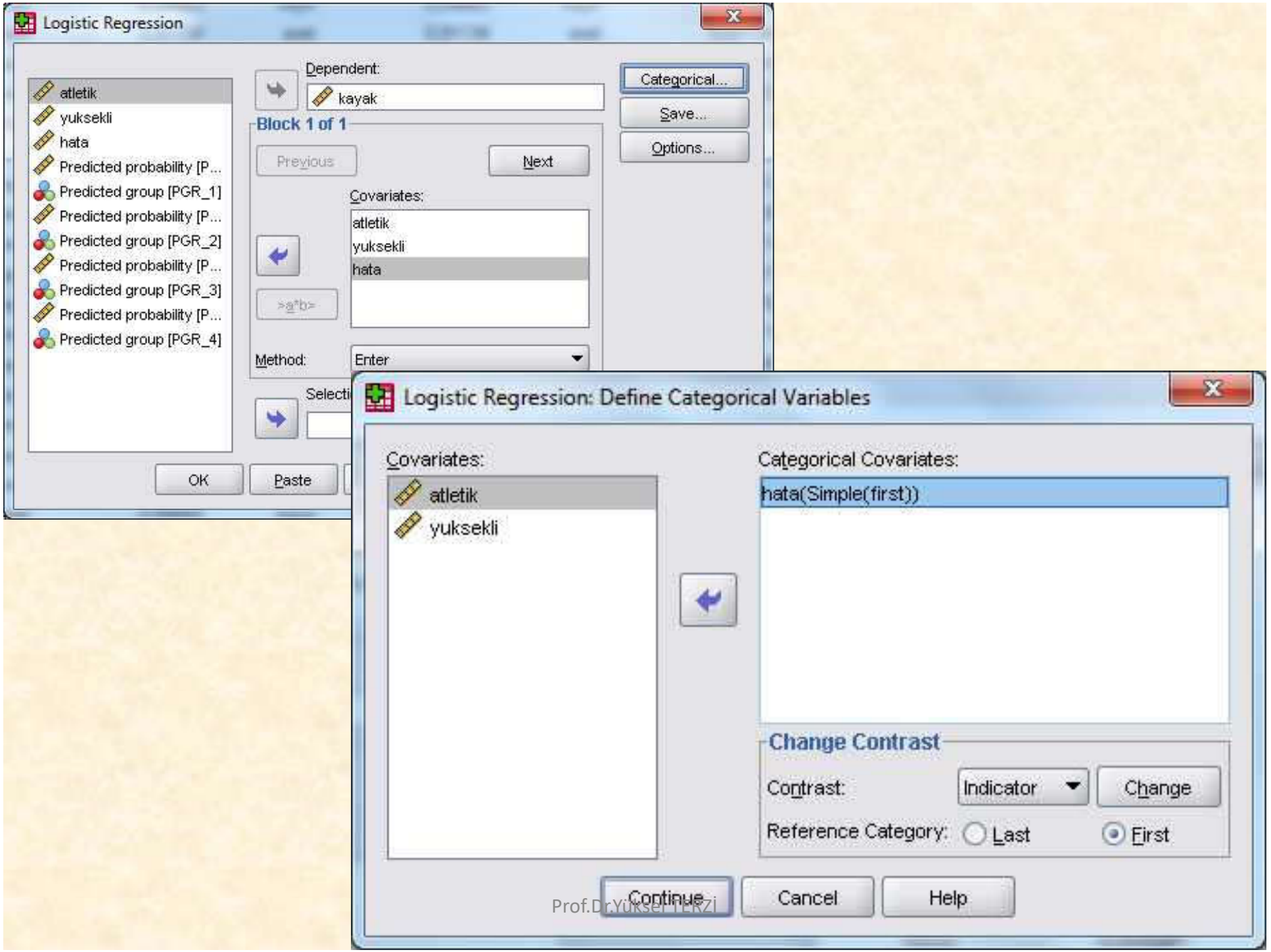
$$OR = Exp(\beta_1) = e^{2,0935} = 8,11$$

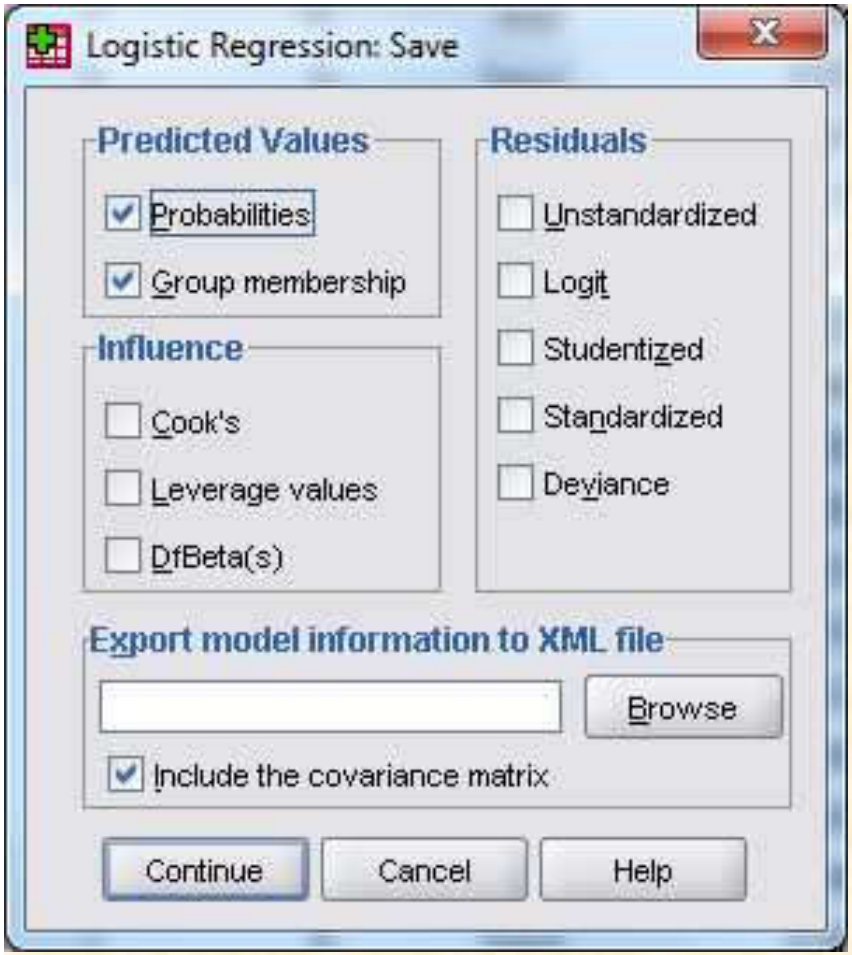
Yaşı ileri olanların (50+) hastalığa yakalanma riski, yaşı <50 olanlara göre 8.11 kat daha fazladır.

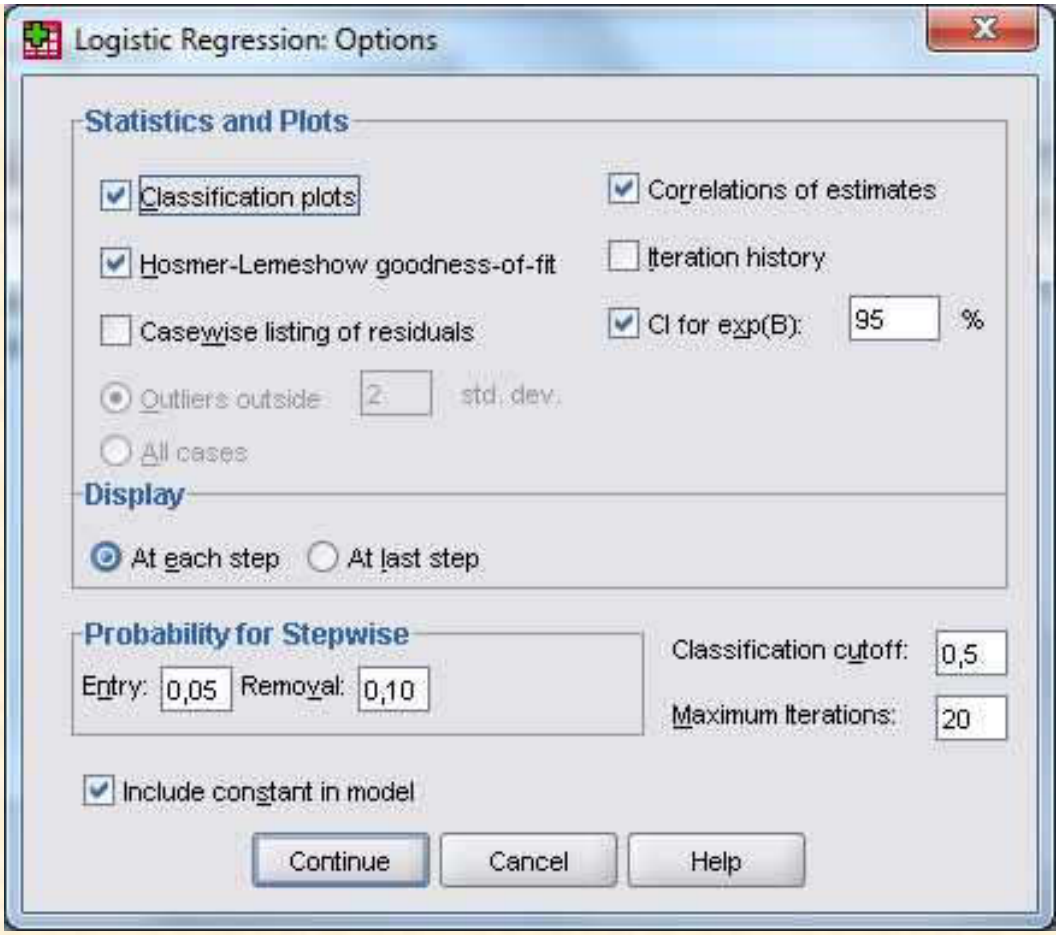
Örnek : **Kayak yapmak isteyenlerle ilgili bir anket yapılıyor. Kayak değişkeni bağımlı değişken alınmış ve evet diyenler (1), hayır diyenler (0) alınmıştır. Tahmin edici değişkenler olarak atletik, yükselti ve hata alınmıştır. Atletik değişkende 0 hiç kabiliyeti olmayanları, 10 ise çok yüksek kabiliyeti göstermektedir. Yükseklikte de yine benzer şekilde kodlama yapılmıştır. Hata değişkeni ise evet (1) ve hayır (0) olarak kodlanmıştır.**

	kayak	atletik	yuksekli	hata
1	1	10	0	1
2	1	4	5	0
3	0	7	10	1
4	1	6	0	0
5	1	6	0	0
6	1	5	0	1
7	1	4	0	0
8	1	3	1	1
9	1	3	0	0
10	0	2	9	1
11	0	2	8	1
12	0	2	6	0
13	0	7	5	0
14	0	7	4	1
15	1	7	3	0
16	0	4	0	0
17	1	9	0	1
18	0	1	6	0
19	1	9	0	1
20	0	1	9	1
21	1	10	0	0
22	0	2	0	1
23	1	9	0	1
24	0	3	0	0






Logistic Regression: Save
Predicted Values
Probabilities
Group membership
Influence
Cook's
Leverage values
DfBeta(s)
Residuals
Unstandardized
Logit
Studentized
Standardized
Deviance
Export model information to XML file
Browse
Include the covariance matrix
Continue
Cancel
Help


Logistic Regression: Options
Statistics and Plots
Classification plots
Hosmer-Lemeshow goodness-of-fit
Casewise listing of residuals
Outliers outside
2
std. dev.
All cases
Correlations of estimates
Iteration history
CI for exp(B):
95
%
Display
At each step
At last step
Probability for Stepwise
Entry:
0,05
Removal:
0,10
Classification cutoff:
0,5
Maximum Iterations:
20
Include constant in model
Continue
Cancel
Help

Block 0: Beginning Block

Iteration History[a,b,c]

Iteration		-2 Log likelihood	Coefficients
			Constant
Step 0	1	55,051	,200
	2	55,051	,201

a. Constant is included in the model.

b. Initial -2 Log Likelihood: 55,051

c. Estimation terminated at iteration number 2 because parameter estimates changed by less than .001

Block 1: Method = Enter

Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients			
			Constant	ATLETIK	YUKSEKLI	HATA(1)
Step 1	1	29,762	-1,697	,385	-,185	,384
	2	25,573	-2,426	,587	-,344	,510
	3	24,480	-2,877	,743	-,481	,459
	4	24,318	-3,090	,830	-,560	,383
	5	24,313	-3,136	,850	-,578	,364
	6	24,313	-3,138	,851	-,579	,363

a. Method: Enter

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 55,051

d. Estimation terminated at iteration number 6 because log-likelihood decreased by less than ,010 percent.

Model uyum iyiliği testi:

Gözlenen değerlerin tahmin edilen değerlerle karşılaştırması için -2logL değerlerine bakılır. İyi model gözlenen sonuçların yüksek ihtimallerini oluşturan modeldir yani -2LL değerinin (sıfıra yakınsa model mükemmeldir) küçük olması gerekir.

Block 0 da sadece sabitin yer aldığı model için -2LL=55.0.51 bulundu.

Block 1 de ise sabit ve tüm bağımsız değişkenlerin yer aldığı modelin -2LL değeri 24.313 bulunmuştur.

Model Chi-Square sadece sabiti olan modelin -2LL değeri ile bütün değişkenleri ihtiva eden modelin -2LL arasındaki farkı verir. Bu istatistik ile sabit hariç, tüm değişkenlerin katsayılarının sıfır olduğu yokluk hipotezi test edilir.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	30,739	3	,000
	Block	30,739	3	,000
	Model	30,739	3	,000

P=0,00<0.05 olduğundan bağımsız değişkenlerin katsayıları sıfırdan farklıdır.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	24,313[a]	,536	,717

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than ,001.

Modelin verileri nasıl temsil ettiği gösteren tahmini olasılıklarla gerçek olasılıklar arasındaki ilişki ve -2logL istatistikleri verilmiştir. Lojistik regresyonda bağımlı değişken ile bağımsız değişkenler arasındaki ilişkinin gücü Cox-Snell R^2 ile Nagelkerle R^2 değerleri ile belirlenir.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	8,355	7	,302

Modelin uyum iyiliğini incelemek amacıyla doğru sınıflandırma yüzdesi ve Hosmer-Lemeshow ölçütleri kullanılır. **Bu istatistik lojistik regresyon modelini genel olarak test eder.**

H_0: Sabit terim dışındaki tüm katsayılar sıfırdır.

Hosmer-Lemeshow istatistiği olabilirlik oran testidir ve modelde bağımsız değişkenlerin olmadığı $-2\log L_0$ istatistiği ile modelde bağımsız değişkenlerin yer aldığı -2logL istatistiği arasındaki fark alınarak hesaplanır. Bu istatistik incelenen modelin parametre sayısı ile sabit terimli modelin parametreleri arasındaki fark bir serbestlik dereceli ki-kare dağılımına uyar. Modelin anlamlı olması arzu edilir ($p<0,05$).

Classification Table[a]

Observed			Predicted		
			KAYAK		Percentage Correct
			hayýr	evet	
Step 1	KAYAK	hayýr	15	3	83,3
		evet	4	18	81,8
	Overall Percentage				82,5

a. The cut value is ,500

Classification Table de görüldüğü gibi sadece 7 denek yanlış sınıflanmıştır. Tablodan kayak yapmama kararı verenlerin %83.3'ü, yapacak olanların ise %81.8'i doğru tahmin edilmiştir. Genel olarak doğru sınıflama oranı %82,5 olmuştur.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	atletik	,851	,311	7,501	1	,006	2,342	1,274	4,305
	yuksekli	-,579	,309	3,508	1	,061	,560	,306	1,027
	hata(1)	-,363	1,038	,122	1	,727	,696	,091	5,319
	Constant	-2,956	1,237	5,713	1	,017	,052		

a. Variable(s) entered on step 1: atletik, yuksekli, hata.

Yukarıdaki tabloda B değerleri katsayılar olup, deneğin bir işi yada diğerini yapma ihtimalini belirlemede kullanılır. B sütunundaki işaretler ilişkinin yönünü gösterir.

Wald ist.=$(B/S.E.)^2$ ile bulunur. Burada Atletik katsayı önemli (P=0,006<0.05), diğerleri ise önemsiz bulunmuştur.

Exp(B) odds oranlarıdır. Yani kayak yapmaya karar verenlerin ihtimalinin, kayak istemeyenlerin ihtimaline oranıdır. Mesela atletik kabiliyet bir birim artarsa In odds 2,342 kat artar.

1.kişi

$$= \frac{1}{1+\exp(-(-2{,}956+0{,}851X_1-0{,}579X_2-0{,}363X_3))}$$

$$= \frac{1}{1+\exp(2{,}956-0{,}851(10)+0{,}579(0)+0{,}363(1))} = 0{,}9945$$

İhtimal 0,5'ten büyükse bu kişinin kayak yapacağı tahmin edilir. 0,5'ten küçük ise kayak yapmayacağı tahmin edilir.

```
             Step number: 1

             Observed Groups and Predicted Probabilities

       8 +                                                                                                    +
         |                                                                                                    |
         |                                                                                                    |
F        |                                                                                                    |
R      6 +                                                                                                  e +
E        |                                                                                                  e |
Q        |                                                                                                  ee|
U        |                                                                                                  ee|
E      4 +h                  h                                                                              ee+
N        |h                  h                                                                              ee|
C        |h                  h                        e                                                     ee|
Y        |h                  h                        e                                                     ee|
       2 +h                  h                        h                    e                         e      ee+
         |h                  h                        h                    e                         e      ee|
         |hh  h h  e    h h  h   e                  e h            h    h  h         e    e          e e   eee|
         |hh  h h  e    h h  h   e                  e h            h    h  h         e    e          e e   eee|
Predicted ---------+---------+---------+---------+---------+---------+---------+---------+---------+----------
  Prob:   0        ,1        ,2        ,3        ,4        ,5        ,6        ,7        ,8        ,9         1
  Group:  hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhheeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee

          Predicted Probability is of Membership for evet
          The Cut Value is ,50
          Symbols: h - hayır
                   e - evet
          Each Symbol Represents ,5 Cases.
```

E kayak yapanlar, H yapmayanları gösterir. Eğer E'lerin çoğu 0,5'in sağında ve H'lerin büyük kısmı 0.5'in solunda ise doğru sınıflandırma yapılmıştır.

Casewise List[b]

Case	Selected Status[a]	Observed kayak	Predicted	Predicted Group	Temporary Variable Resid	Temporary Variable ZResid
2	S	e**	,094	h	,906	3,107

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2,000 are listed.

2. Gözlemin student artık değeri 2'den büyük, standardize artık değeri ise 3,107 olduğundan ve yanlış sınıflandırma yapıldığından silinmelidir.

Veri sayfasında tahmin edilen değerlerin ihtimalleri pre_1 ve tahmin edilen grup üyeleri pgr_1 ile verilmiştir. Pre_1>0,5 olanlar evet, <0,05 olanlar ise hayır olarak tahmin edilir.

PRE_1	PGR_1	kayak
0,99537	evet	evet
0,09385	hayır	evet
0,04863	hayır	hayır
0,91134	evet	evet
0,91134	evet	evet
0,75330	evet	evet
0,65212	evet	evet
0,23784	hayır	evet
0,44461	hayır	evet
0,00129	hayır	hayır
0,00231	hayır	hayır
0,01047	hayır	hayır
0,57078	evet	hayır
0,62277	evet	hayır
0,80898	evet	evet
0,65212	evet	hayır
0,98923	evet	evet
0,00450	hayır	hayır
0,98923	evet	evet
0,00055	hayır	hayır
0,99677	evet	evet
0,19213	hayır	hayır
0,98923	evet	evet
0,44461	hayır	hayır

kayak	Numeric	8	0	
atletik	Numeric	8	0	
yuksekli	Numeric	8	0	
hata	Numeric	8	0	
PGR_1	Numeric	8	0	Predicted group
PRE_1	Numeric	11	5	Predicted probability
COO_1	Numeric	11	5	Analog of Cook's influence statistics
LEV_1	Numeric	11	5	Leverage value
RES_1	Numeric	11	5	Difference between observed and predicted probabilities
LRE_1	Numeric	11	5	Logit residual
SRE_1	Numeric	11	5	Standard residual
DEV_1	Numeric	11	5	Deviance value
DFB0_1	Numeric	11	5	DFBETA for constant
DFB1_1	Numeric	11	5	DFBETA for yuksekli
DFB2_1	Numeric	11	5	DFBETA for hata
DFB3_1	Numeric	11	5	DFBETA for atletik

PRE_1	COO_1	LEV_1	RES_1	LRE_1	SRE_1	DEV_1	DFB0_1	DFB1_1	DFB2_1	DFB3_1
1,00000	0,00000	0,00007	0,00000	1,00000	0,00145	0,00145	-0,00003	-0,00001	0,00001	0,00001
0,00001	0,00000	0,00036	-0,00001	-1,00001	-0,00347	-0,00347	-0,00013	-0,00006	0,00005	0,00004
0,99188	0,00054	0,06221	0,00812	1,00818	0,13184	0,12767	-0,07088	-0,02715	0,02065	0,02296
0,99188	0,00054	0,06221	0,00812	1,00818	0,13184	0,12767	-0,07088	-0,02715	0,02065	0,02296
0,98561	0,00209	0,12507	0,01439	1,01460	0,18200	0,17024	-0,15599	-0,05628	0,06881	0,04644
0,72930	0,11676	0,23929	0,27070	1,37117	0,91100	0,79457	-0,42943	-0,28936	-0,00182	0,21519
0,16888	0,74014	0,13073	0,83112	5,92148	2,02290	1,88605	1,27204	0,39373	0,33002	-0,36842

$$Logit\ Hata_i = LRE = \frac{e_3}{P_3(1-P_3)} = \frac{0{,}00812}{0{,}99188(1-0{,}99188)} = 1{,}00819$$

PRE_1	COO_1	LEV_1	RES_1	LRE_1	SRE_1	DEV_1	DFB0_1	DFB1_1	DFB2_1	DFB3_1
1,00000	0,00000	0,00007	0,00000	1,00000	0,00145	0,00145	-0,00003	-0,00001	0,00001	0,00001
0,00001	0,00000	0,00036	-0,00001	-1,00001	-0,00347	-0,00347	-0,00013	-0,00006	0,00005	0,00004
0,99188	0,00054	0,06221	0,00812	1,00818	0,13184	0,12767	-0,07088	-0,02715	0,02065	0,02296
0,99188	0,00054	0,06221	0,00812	1,00818	0,13184	0,12767	-0,07088	-0,02715	0,02065	0,02296
0,98561	0,00209	0,12507	0,01439	1,01460	0,18200	0,17024	-0,15599	-0,05628	0,06881	0,04644
0,72930	0,11676	0,23929	0,27070	1,37117	0,91100	0,79457	-0,42943	-0,28936	-0,00182	0,21519
0,16888	0,74014	0,13073	0,83112	5,92148	2,02290	1,88605	1,27204	0,39373	0,33002	-0,36842

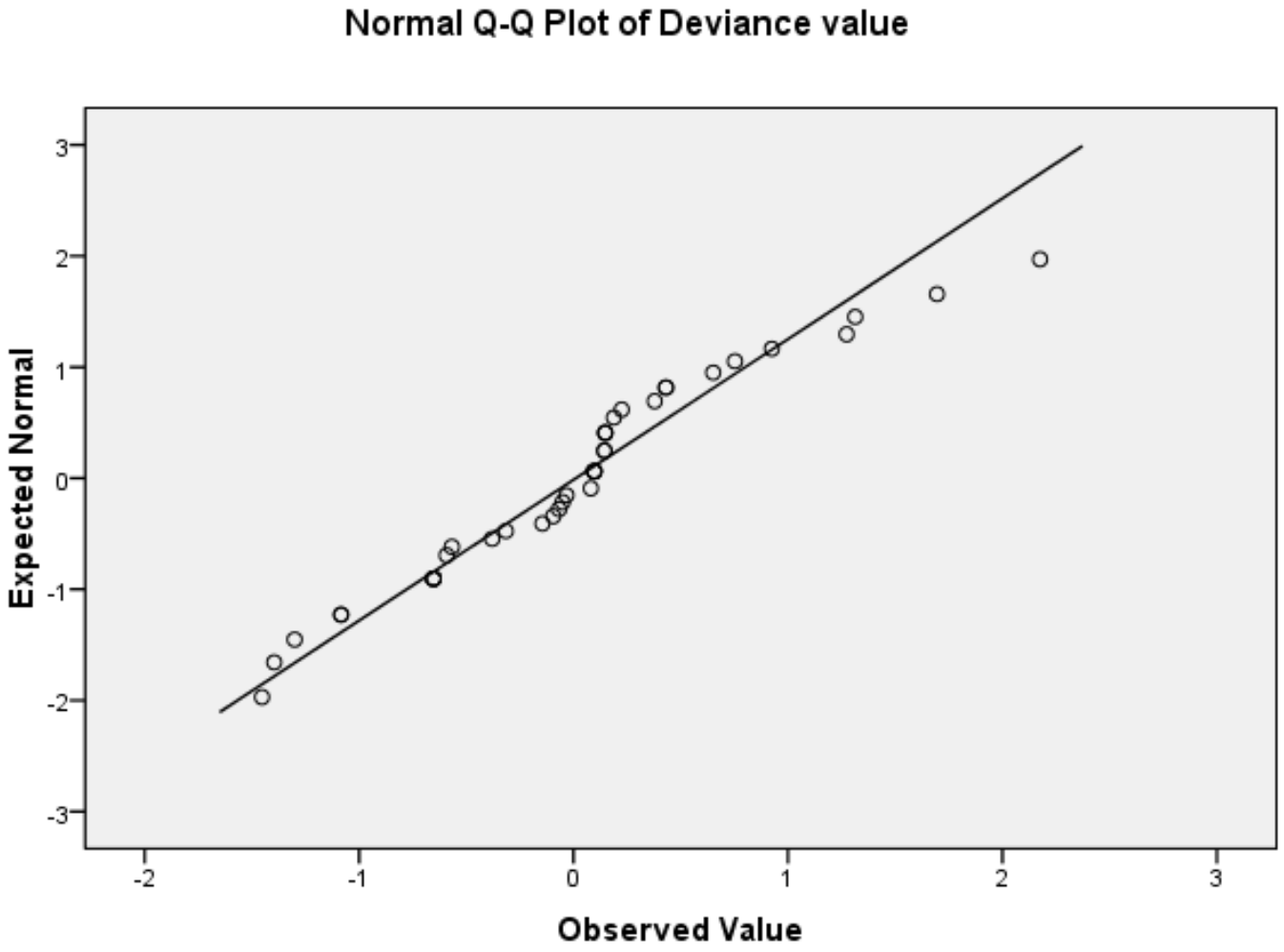
3. Gözlem için

Standardize Hata(SRE):

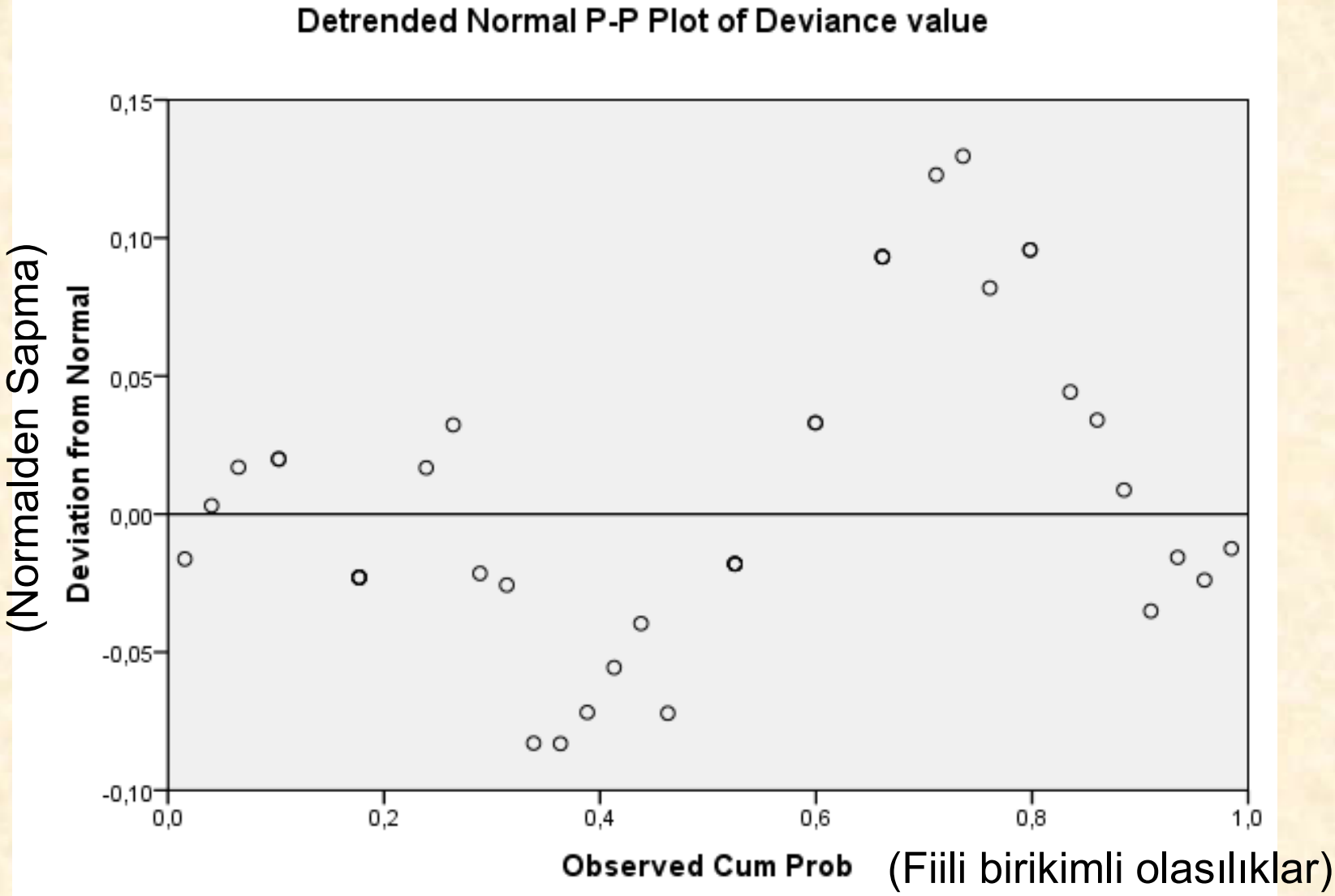
$$Z_3 = \frac{e_3}{\sqrt{P_3(1-P_3)}} = \frac{0,00812}{\sqrt{0,99188(1-0,99188)}} = 0,13$$

$$Sapma(Deviance) = \sqrt{-2\ln(P_3)} = \sqrt{-2\ln(0,99188)} = 0,127$$

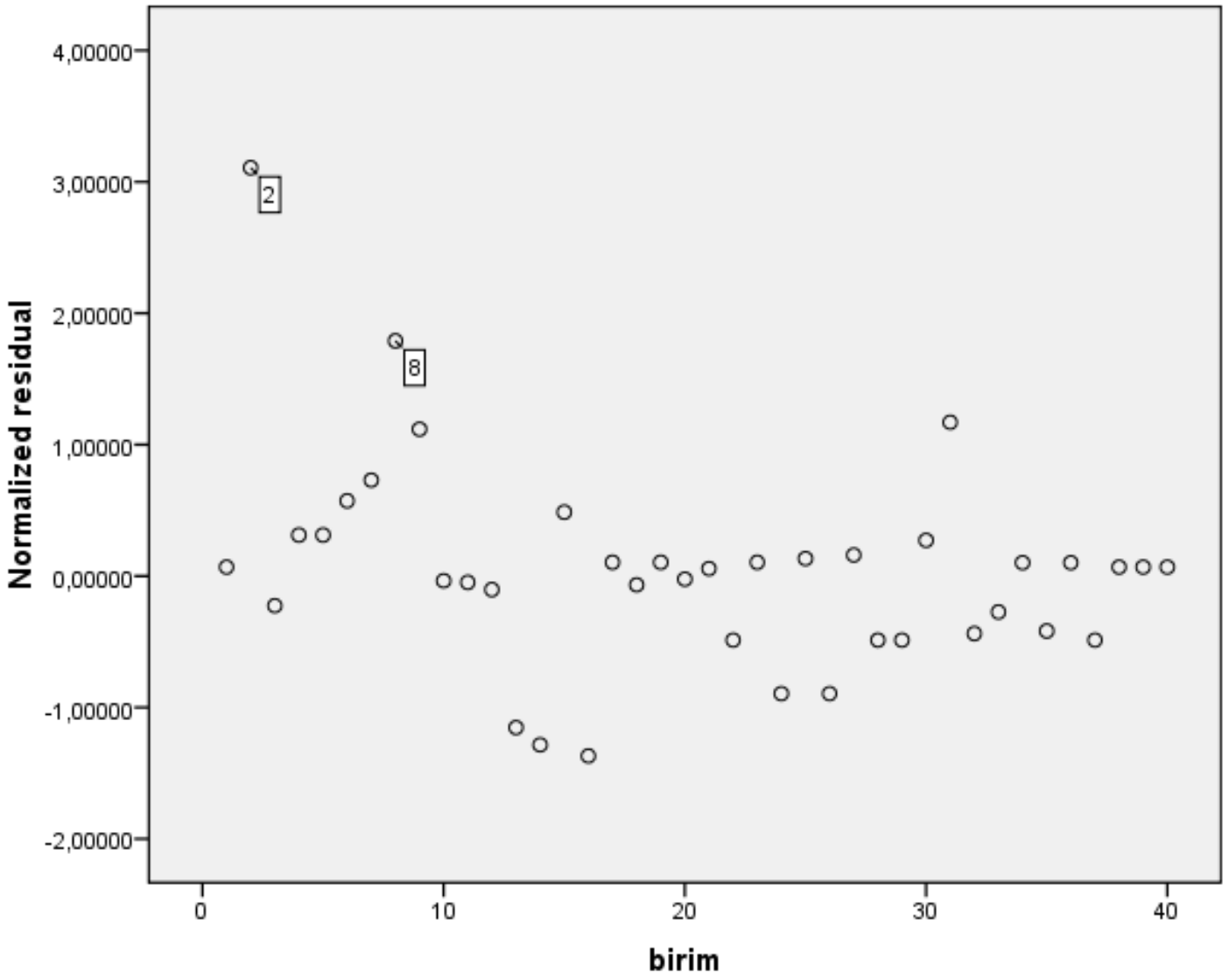
$$Cook_3 = Z_3^2\left(\frac{h_3}{1-h_3}\right) = 0,13184^2\left(\frac{0,06221}{1-0,06221}\right) = 0,001$$



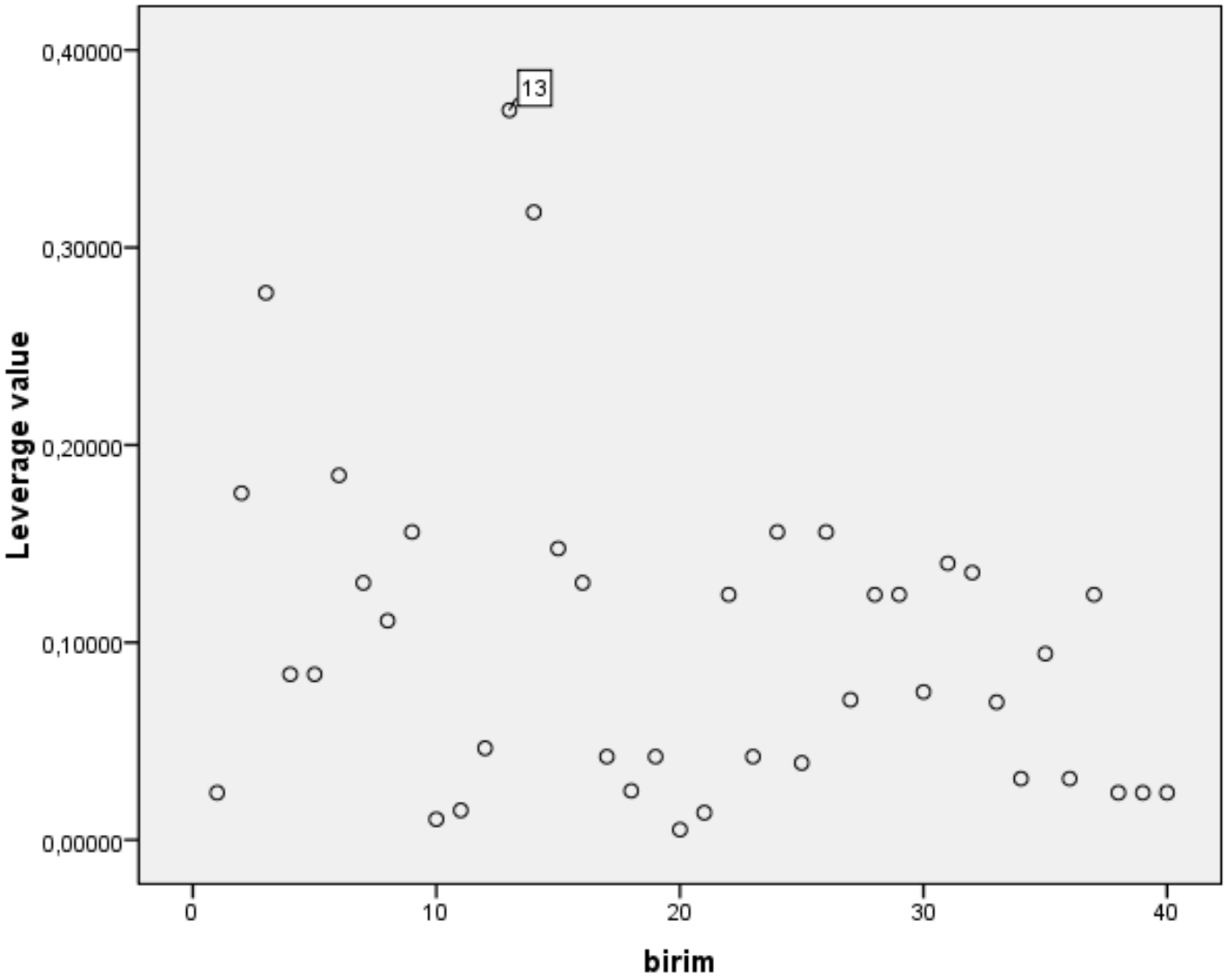
Sapma (Deviance) değerlerin normallik varsayımı için Q-Q grafiği



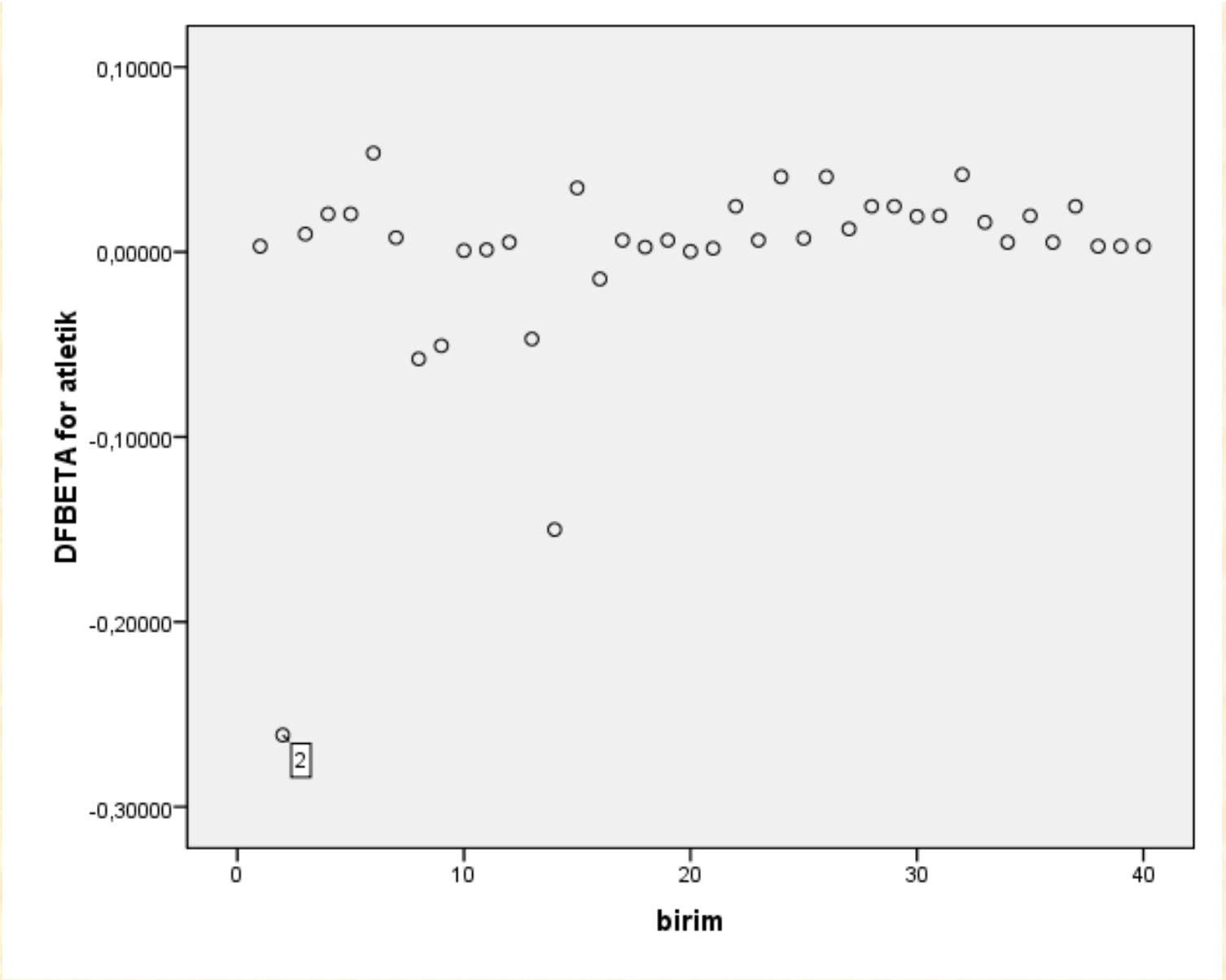
Sapma (Deviance) değerlerin normallik varsayımı için trendsiz P-P grafiği



Birim sırasına göre standardize artıkların serpme diyagramı.
2. Gözlemin standardize artık değeri 3'den büyük olduğundan uç değerdir (bağımsız değişkenlerdeki (Xi) kuşkulu gözlemdir).



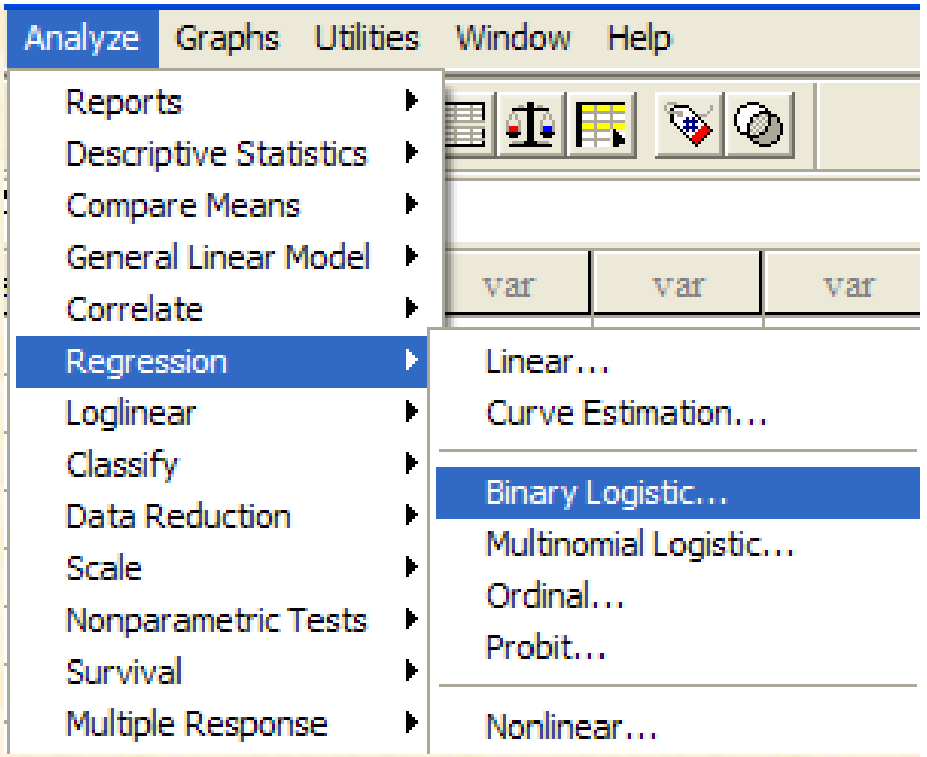
Birim sırasına göre uzaklık (leverage) değerlerinin serpme diyagramı 13. ve 14. gözlemler etkili gözlemdir (silindiğinde hesaplamalarda önemli ölçüde değişikliğe sebep olan gözlemler).

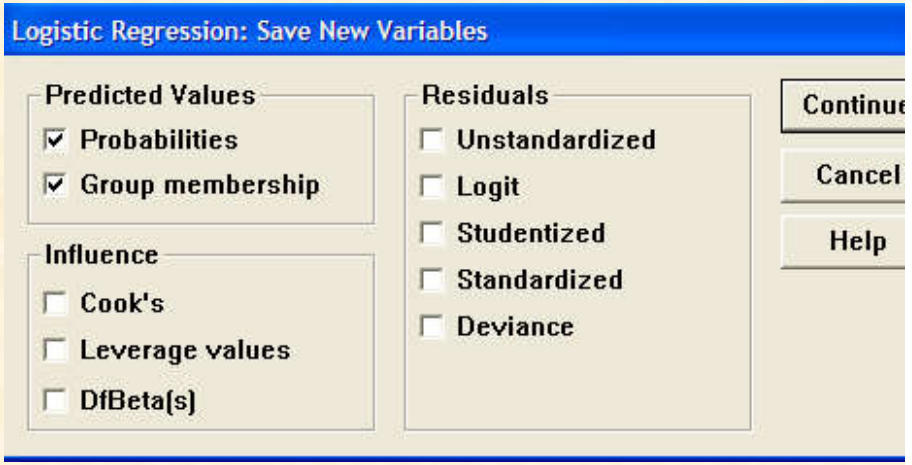


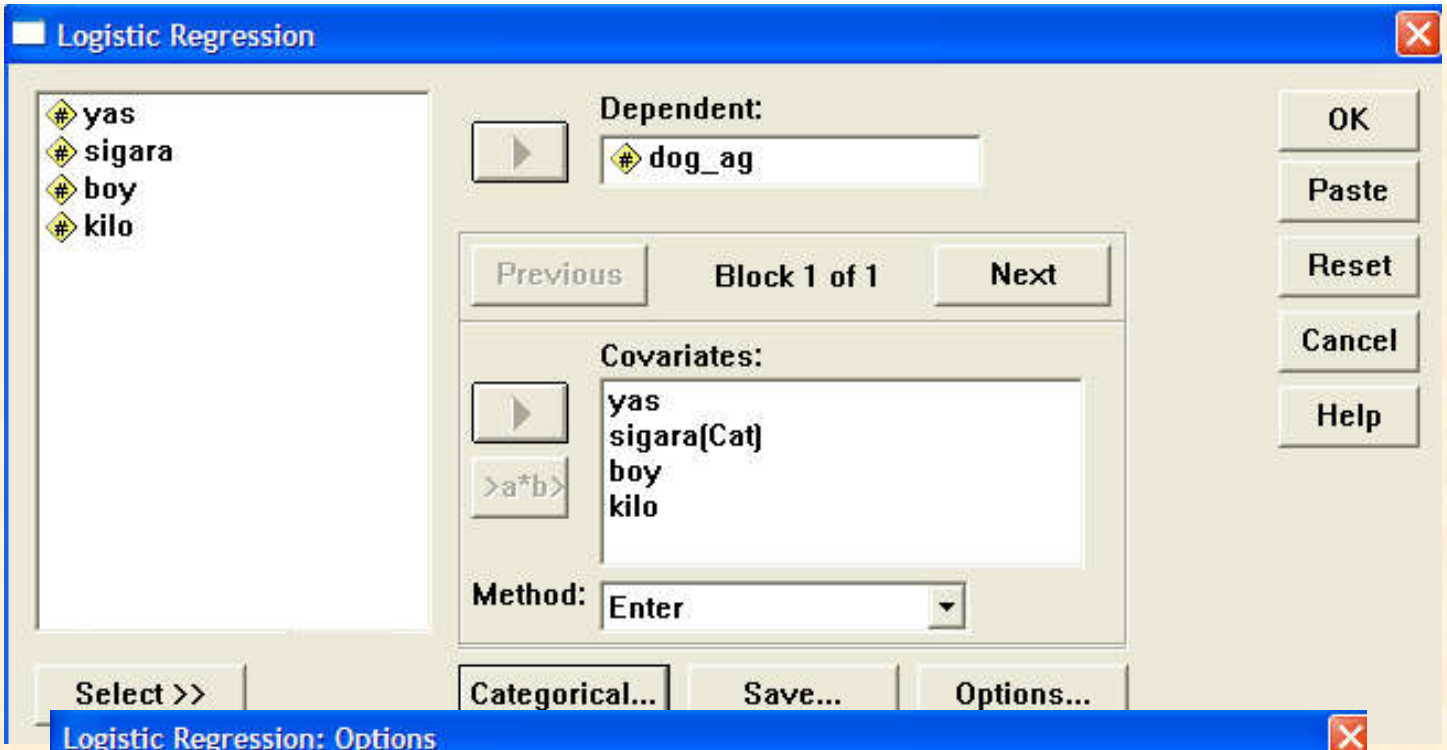
Atletik değişkeni için DfBeta değerlerinin serpme diyagramı

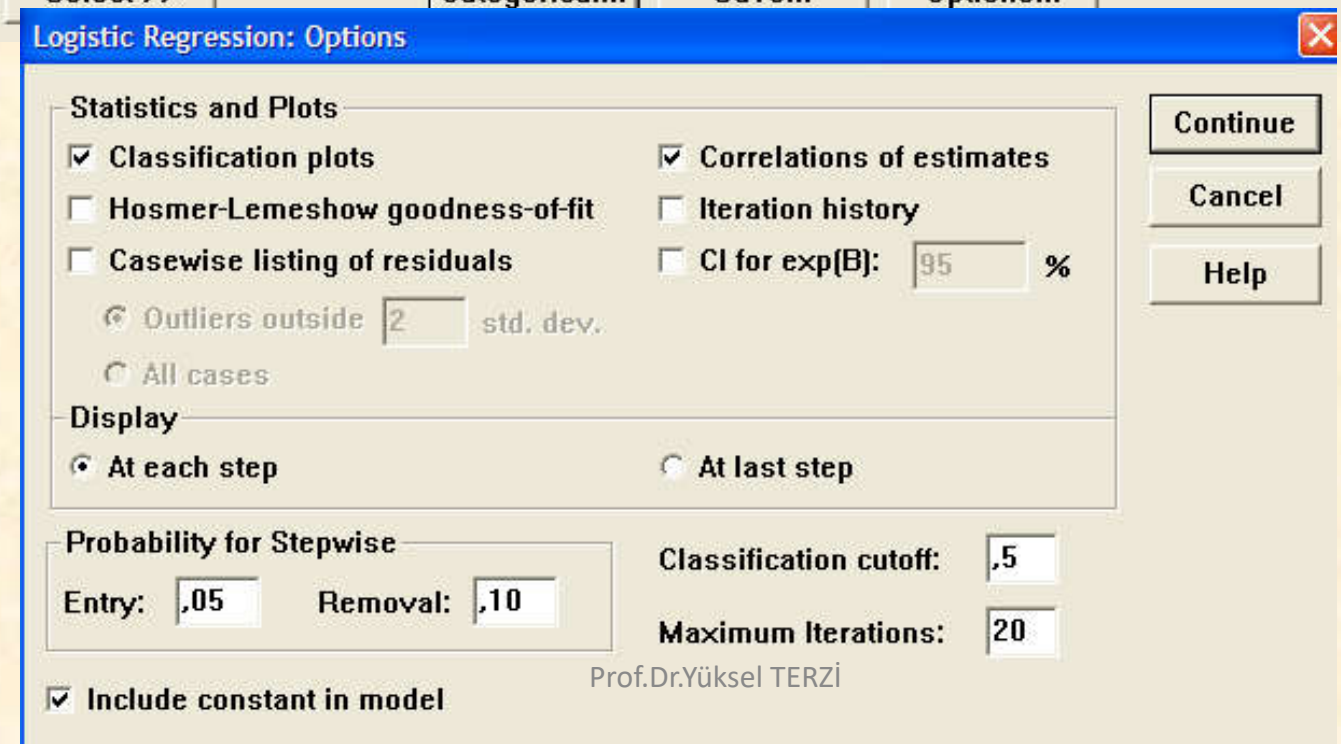
Örnek.

	dog_ag	yas	sigara	boy	kilo
1	1,00	33,00	1,00	168,00	58,78
2	1,00	19,00	1,00	170,00	57,79
3	2,00	29,00	,00	164,00	56,80
4	2,00	27,00	,00	151,00	41,42
5	1,00	30,00	1,00	166,00	65,15
6	2,00	18,00	,00	168,00	55,83
7	1,00	21,00	1,00	157,00	56,72
8	2,00	13,00	1,00	166,00	55,09
9	1,00	33,00	,00	170,00	60,84
10	1,00	28,00	1,00	157,00	60,79
11	2,00	32,00	1,00	165,00	66,68
12	2,00	28,00	,00	157,00	49,58
13	2,00	23,00	,00	162,00	57,15
14	1,00	32,00	1,00	165,00	58,49
15	2,00	28,00	,00	177,00	78,48
16	1,00	24,00	1,00	170,00	62,59
17	1,00	28,00	,00	172,00	61,98
18	2,00	24,00	,00	159,00	66,21
19	2,00	24,00	,00	155,00	58,47
20	2,00	34,00	1,00	164,00	66,79
21	2,00	24,00	,00	165,00	51,70
22	1,00	30,00	1,00	166,00	59,11
23	2,00	30,00	1,00	164,00	55,02
24	2,00	26,00	1,00	161,00	55,58
25	1,00	28,00	,00	161,00	67,98
26	1,00	11,00	,00	162,00	64,88
27	2,00	24,00	,00	171,00	61,20
28	2,00	28,00	,00	163,00	60,92
29	1,00	26,00	1,00	165,00	56,74
30	1,00	34,00	,00	160,00	56,90






Logistic Regression
yas
sigara
boy
kilo
Dependent:
dog_ag
OK
Paste
Reset
Cancel
Help
Previous
Block 1 of 1
Next
Covariates:
yas
sigara(Cat)
boy
kilo
>a*b>
Method:
Enter
Select >>
Categorical...
Save...
Options...


Logistic Regression: Options
Statistics and Plots
Classification plots
Hosmer-Lemeshow goodness-of-fit
Casewise listing of residuals
Outliers outside
2
std. dev.
All cases
Correlations of estimates
Iteration history
CI for exp(B):
95
%
Continue
Cancel
Help
Display
At each step
At last step
Probability for Stepwise
Entry:
,05
Removal:
,10
Classification cutoff:
,5
Maximum Iterations:
20
Include constant in model

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	YAS	-,019	,071	,074	1	,786	,981
	SIGARA(1)	1,362	,797	2,921	1	,087	3,902
	BOY	-,027	,084	,103	1	,748	,973
	KILO	-,042	,074	,326	1	,568	,958
	Constant	6,883	12,501	,303	1	,582	975,339

a. Variable(s) entered on step 1: YAS, SIGARA, BOY, KILO.

Exp(B)=OR>1 ise ve katsayısı önemli ise o değişkenler önemli bir risk faktörü taşımaktadır. Burada tüm değişkenlerin doğum ağırlığı üzerinde etkisi yoktur. Ancak sigara içmenin içmemeye göre düşük doğum ağırlıklı bebek doğumuna 3.9 kat daha fazla olduğu görülmektedir.

$$P(Y) = \frac{1}{1+\exp(-6{,}88+0{,}19Yas-1{,}36Sigara+0{,}027Boy+0{,}04Kilo)}$$

30 yaşında, 170 cm boyunda, 70 kg ve sigara içen bir annenin düşük doğum yapıp yapmayacağını tahmin edelim. P(Y)<0.5 olduğunda düşük doğum, P(Y)>=0.5 ise normal doğum olması beklenir.

$$P(Y) = \frac{1}{1+\exp(-6{,}88+0{,}19*30-1{,}36*1+0{,}027*170+0{,}04*70)} = 0.99$$

2. Sıralı Lojistik Regresyon (Ordinal Logistic Regression):

Cevap değişkenin sıralı ölçekli olduğu durumlarda uygulanan bir yöntem olup, en az üç kategorisi olması gerekir(hafif-orta-ağır gibi). Açıklayıcı değişkenler faktör yada ortak değişkelerdir (covariate).

Parametre tahminleri yinelemeli-ağırlıklı en küçük kareler yöntemine (iterative-reweighted least square method) göre en büyük benzerlik parametre tahminleri yapar. **Kategoriler birbirine paraleldir varsayımı kullanılır. En büyük değere sahip cevap referans alınarak bu referansa göre lojit modeller türetilerek analiz yapılır.**

Sıralı kategorik bağımsız değişken düzeylerinin her bir kombinasyonu için elde edilen hata terimlerinden rassal olarak biri seçileceği için, hata terimleri normal dağılmayacaktır. Böylece klasik regresyon modelinin "hata terimleri normal dağılımlıdır" varsayımı ihlal edilmiş olur.

Bağımlı değişken sürekli olmadığı için en küçük kareler tekniği anlamsız kestirimler verecektir.

Nominal bağımlı değişkenler için kodlama tamamen gelişi güzel yapılır ve sıralı bağımlı değişken için sabit bir dönüşüme kadar kodlama keyfi olacaktır. Ancak kaydedilen bağımlı değişken çok farklı sonuçlar verecektir.

Sıralı lojistik regresyon modellerinin elde edilmesinde beş temel bağlantı fonksiyonu (link function) kullanılmaktadır. En sık kullanılan fonksiyonlar ise logit, probit ve log log fonksiyonlarıdır.

Sıralı lojistik regresyon modelinin başlıca özellikleri:

1. İlgili bağımlı değişken, gözlemlenmemiş sürekli gizli (latent) bir değişkenden tekrar düzenlenebilir sıralı ve gruplanmış bir kategorik değişkendir. Ancak sıralı değişkenin kategorilerinin eşit aralıklarla ayrılıp ayrılmadığı kesin değildir.
2. **Sıralı lojistik regresyon analizi, normallik ve sabit varyans varsayımını gerektirmeden, açıklayıcı değişkenlerin sıralı kategorik değişken üzerindeki etkilerini açıklamak için bağlantı fonksiyonu kullanır.**
3. Regresyon katsayısının değeri sıralı kategorik değişkeninin kategorilerine bağlı olmadığından dolayı sıralı lojistik regresyon modeli, açıklayıcı değişken ile sıralı kategorik değişken arasındaki ilişkinin kategoriden bağımsız olduğunu varsayar. Bağlantı fonksiyonu kullanılarak tahmin edilen regresyon katsayıları her bir kesme noktasında (eşik değerinde) aynıdır.

Sıralı lojistik regresyon modeli, McCullagh (1980) tarafından geliştirilmiş bir modeldir. Model, gözlemlenebilir bir kategorik değişkenin altında gözlemlenemeyen bir gizli değişkenin olduğu varsayımına dayandırılmaktadır.

Sıralı Lojistik Regresyon Modeli:

$$link(\gamma_j) = \tau_j - \sum \beta'_k x_k$$

Sıralı bağımlı değişkenin kategori sayısı üç ve üçten büyük olduğunda:

$$link(\gamma_j) = \frac{\tau_j - [\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k]}{\exp(\theta_0 + \theta_1 z_1 + \theta_2 z_2 + \ldots + \theta_l z_l)}$$

Burada, γ_j, j. kategori için birikimli olasılık değeri, τ_j j. kategorinin eşik değeri, $\beta_1 \ldots . \beta_k$ regresyon katsayıları, yer parametreleri için $x_1 \ldots . x_k$ açıklayıcı değişkenler ve k açıklayıcı değişken sayısıdır. β *ve* θ bilinmeyen yer ve ölçek parametreleri vektörüdür. Ayrıca τ_j bilinmeyen kesme noktaları vektörü ve z_l ölçek parametreleri için açıklayıcı değişkenlerdir.

Logit Bağlantı (Link) Fonksiyonu

Logit fonksiyonunu kullanılması durumunda elde edilen modelde hata terimlerinin lojistik dağıldığı varsayıma dayandırıldığı daha önce belirtilmişti. Logit fonksiyon, kategorik bağımlı değişkenin en yüksek kategorisini referans alarak logit modeller üretmektedir. Örneğin üç kategoriye sahip bir bağımlı değişken için üçüncü kategori referans alınarak iki tane logit modeli üretilir. İki kategori için paralel eğriler varsayımı test edilir. Kategorilerin sırasını göz önüne alan sıralı logit modeli için birikimli olasılık değeri

$$\log it(p_1) = \ln\left(\frac{p_1}{1-p_1}\right) = \tau_1 - \beta' x$$

$$\log it(p_1 + p_2) = \ln\left(\frac{p_1 + p_2}{1-p_1-p_2}\right) = \tau_2 - \beta' x$$

$$\log it(p_1 + p_2 + \ldots + p_k) = \ln\left(\frac{p_1 + p_2 + \ldots + p_k}{1-p_1-p_2-\ldots-p_k}\right) = \tau_k - \beta' x$$

Her bir kategori için birikimli olasılık değeri:

$$p_1 = \Pr(Y = 1) = \frac{\exp(\tau_1 - \beta' x)}{1 + \exp(\tau_1 - \beta' x)}$$

$$p_1 + p_2 = \Pr(Y \leq 2) = \frac{\exp(\tau_2 - \beta' x)}{1 + \exp(\tau_2 - \beta' x)}$$

$$p_1 + \ldots + p_k = \Pr(Y \leq k) = \frac{\exp(\tau_k - \beta' x)}{1 + \exp(\tau_k - \beta' x)}$$

Bağlantı (link) fonksiyonu, sıralı lojistik regresyon modelinin yapılandırılmasında kullanılan olasılık fonksiyonudur. Sıralı lojistik regresyon analizinde beş farklı bağlantı fonksiyonu kullanılmaktadır.

Fonksiyon	**Gösterim**	**Uygulama Alanı**
Logit	$\ln\left(\frac{\gamma}{1-\gamma}\right)$	Tüm kategorilerin olasılık değeri eşit ise kullanılır.
Tamamlayıcı Log Log	$\ln(-\ln(1-\gamma))$	Yüksek kategorilerde olasılık değeri daha yüksek ise kullanılır.
Negatif Log Log	$-\ln(-\ln(\gamma))$	Düşük kategorilerde olasılık değeri daha yüksek ise kullanılır.
Probit (normal fonk. tersi)	$\phi^{-1}(\gamma)$	Normal dağılmış gizli değişken söz konusu olduğunda kullanılır.
Cauchit	$\tan(\pi(\gamma-0,5))$	Birçok uç değerin olduğu kategori varsa kullanılır.

Burada γ, bir olayın meydana gelme olasılığıdır.

Sıralı logit model için, u hata terimi 0 ortalama ve $\pi^2/3$ varyansı ile lojistik dağılır.

Olasılık yoğunluk fonksiyonu:

$$f(u) = \frac{\exp(u)}{[1+\exp(u)]^2}$$

Birikimli Dağılım Fonksiyonu:

$$F(u) = \frac{\exp(u)}{1+\exp(u)}$$

Sıralı Probit model için, *u* hata terimi 0 ortalama ve 1 varyans ile normal dağılır.

Olasılık yoğunluk fonksiyonu:

$$f(u)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right)$$

Birikimli Dağılım Fonksiyonu:

$$F(u)=\int_{-\infty}^{u}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right)dt$$

Probit bağlantı fonksiyonu

Probit fonksiyonunun kullanılması durumunda elde edilen modelde **hata**

terimlerinin normal dağılımlı olduğu varsayılır. Probit model gizli bir regresyon denklemi üzerine kurulur. Sıralı lojistik regresyon modelinde, kategorik bağımlı değişkenin altında gizli bir bağımlı değişken olabilir. Sıralı probit modelinde gözlenen ve gizli bağımlı değişken arasındaki ilişkiyi açıklayan kesme noktaları söz konusudur.

Sıralı probit modeli için her bir kategorinin olasılıkları aşağıdaki gibidir:

$$\Pr(Y_i = 1|x_i) = F(\tau_1 - \beta' x_i)$$

$$\Pr(Y_i = 2|x_i) = F(\tau_2 - \beta' x_i) - F(\tau_1 - \beta' x_i)$$

.

.

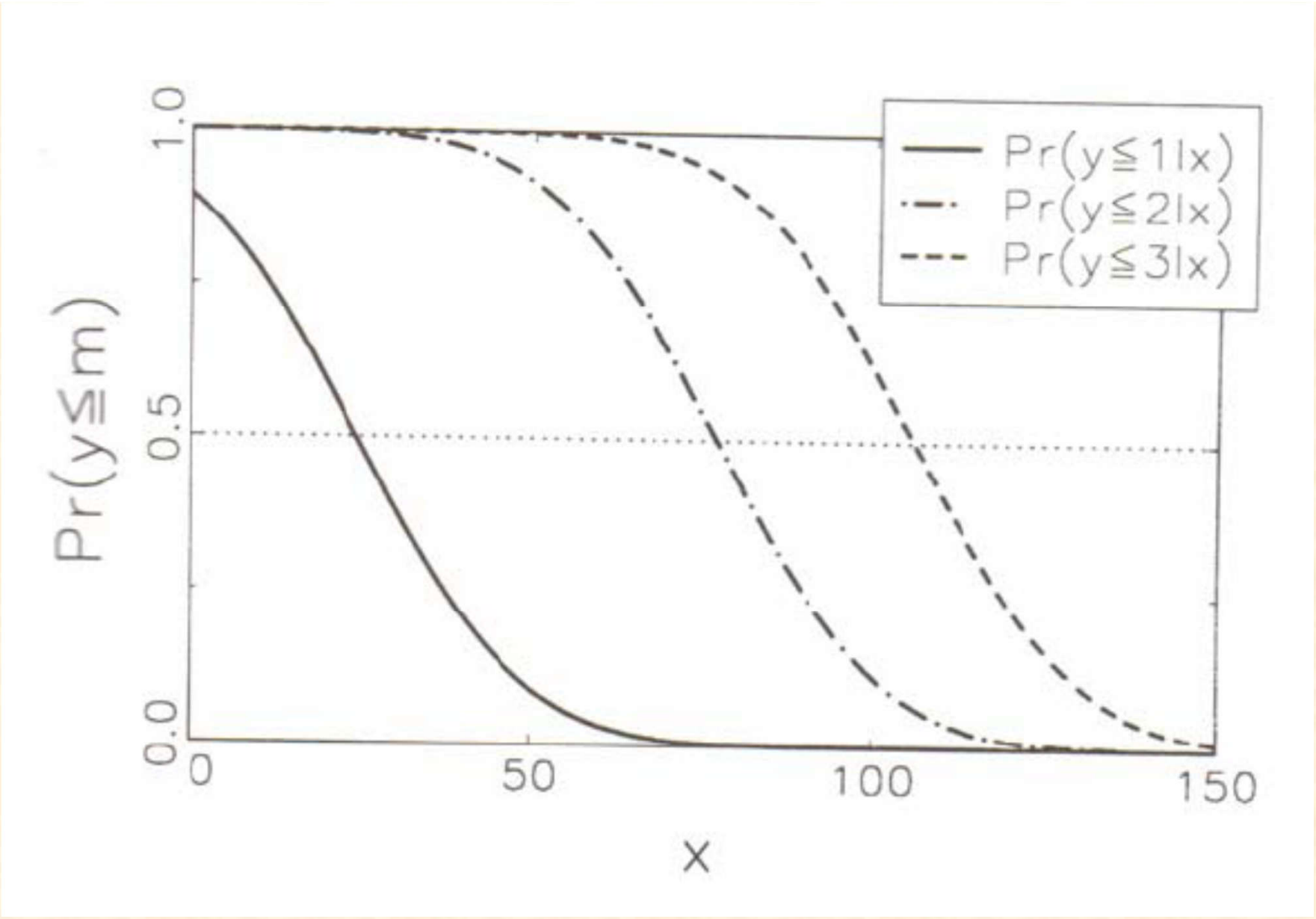
.

$$\Pr(Y_i = J|x_i) = 1 - F(\tau_{J-1} - \beta' x_i)$$

Sıralı Lojistik Regresyon Modelinin Paralel Eğriler Varsayımı

Paralel eğriler varsayımı, **belirlenen regresyon katsayılarının sıralı kategorik değişkenin tüm kategorilerinde eşit olduğunu varsayar**.

Model yapısı genelde çeşitli bağlantı fonksiyonlarını gerektirir. Çünkü bağlantı fonksiyonları sıralı lojistik regresyon modellerini güçlü bir paralel eğriler varsayımı altında oluşturmak için kullanılmaktadır. Bir aday modelin belirlenmesi için temelde, paralel eğriler varsayımının sağlanması gerekir.



Paralel eğriler varsayımı

Sıralı lojistik regresyon modelinden elde edilen bilginin doğruluğu ve güvenirliliği için paralel eğriler varsayımının kesinlikle sağlanması gerekir. Eğer bu varsayım karşılanmazsa elde edilen tüm sonuçlar anlamsız ve yanlış olacaktır. Bu nedenle modelin varsayımı mutlaka test edilmelidir.

Paralel eğriler varsayımının geçerliliğini kontrol etmek için Wald ki-kare testi, olabilirlik oran testi gibi testler kullanılmaktadır

H_0: İlişkili regresyon katsayıları, bağımlı değişkeninin tüm kategorilerinde aynıdır.

H_1: İlişkili regresyon katsayıları, bağımlı değişkeninin tüm düzeylerinde farklıdır.

Sıfır hipotezi reddedilir ise paralel eğriler varsayımı sağlanamaz.

Pseudo R-Square (Sözde R^2):

Modelin uygunluğunu test etmek için kullanılabilecek bir ölçüt de sözde R^2'dir. Bu istatistik çoklu belirlilik katsayısı ile hemen hemen aynıdır.

$$R^2 = 1 - (\ln L / \ln L_0)$$

L_0 : Sadece sabit terimin (β_0) yer aldığı modelin en çok olabilirlik değeridir.

L : Tahmin edilen tüm parametrelerin yer aldığı modelin en çok olabilirlik değeridir.

R^2, değeri verideki belirsizliğin model tarafından açıklanabilen oranını göstermektedir. L=1 olduğunda, lnL=0 ve R^2=1 olur. Bu da ele alınan bağımsız değişkenler tarafından bağımlı değişkendeki değişimin tamamının açıklandığının ve modelin mükemmel olduğunun bir göstergesidir (Ayhan, 2006).

Bağımlı ve bağımsız değişkenler arasındaki ilişkinin gücünün ölçülmesinde kullanılan iki sözde R^2 istatistiği daha vardır. Bu istatistikler, Cox ve Snell R^2 istatistiği ve Nagelkerke R^2 istatistiğidir.

$$R^2_{CS} = 1 - \left(\ln L_0 / \ln L\right)^{\frac{2}{n}}$$

$$R^2_N = \frac{R^2_{CS}}{1 - L_0^{\frac{2}{n}}}$$

Model Parametrelerinin Yorumu

Sıralı lojistik regresyon modeli üç farklı şekilde yorumlanabilir (Long, 1997; Tansel ve Güngör, 2004).

* Standartlaştırılmış katsayılara göre (partial change in y*)
* Tahmin edilen olasılıklara göre (partial change in predicted probabilities)

Standartlaştırılmış katsayıya göre yorum:

Diğer tüm değişkenler sabit tutulduğunda, x_k'daki bir birimlik artış, y^*'da β_k birim kadar değişime neden olmaktadır.

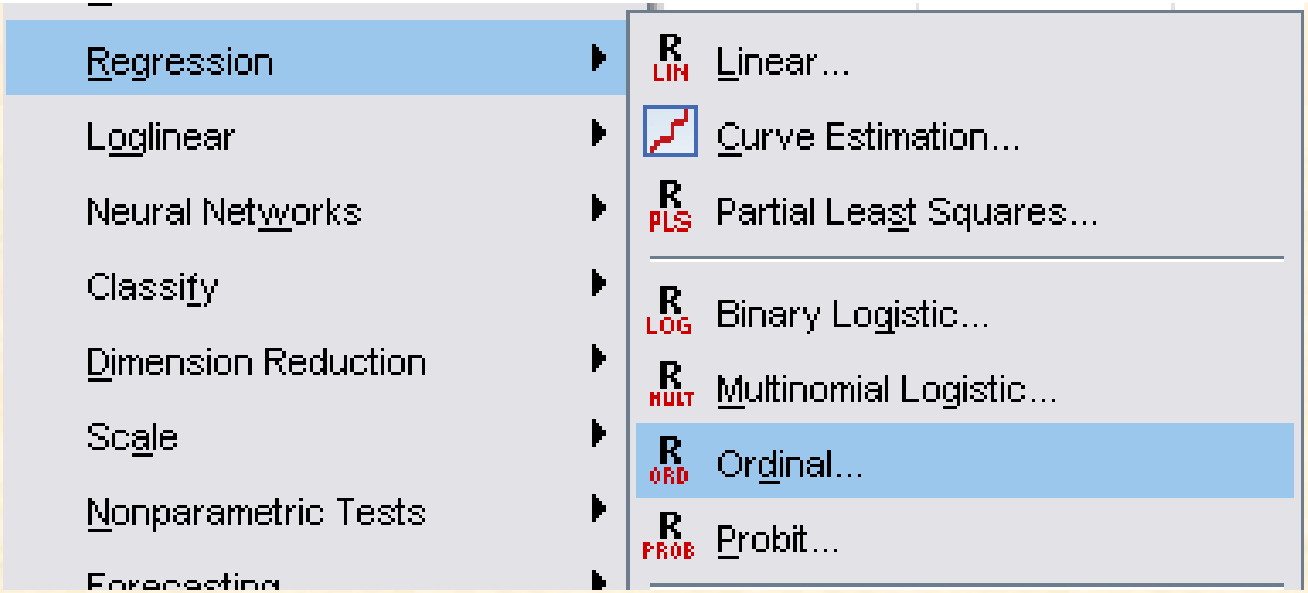
Tahmin edilen olasılıklara göre yorum:

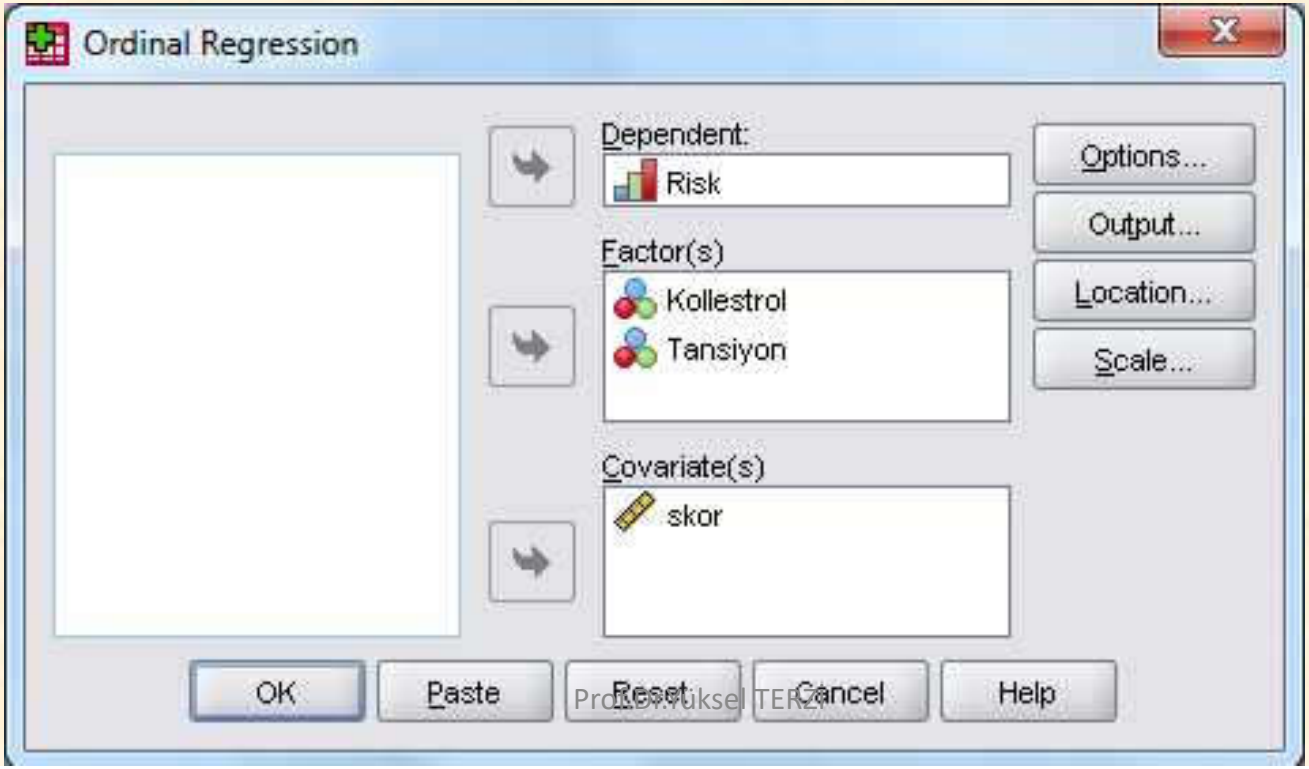
Bağımsız değişkendeki bir birimlik değişimin bağımlı değişken kategorisinde gerçekleşme olasılığına katkısıdır.

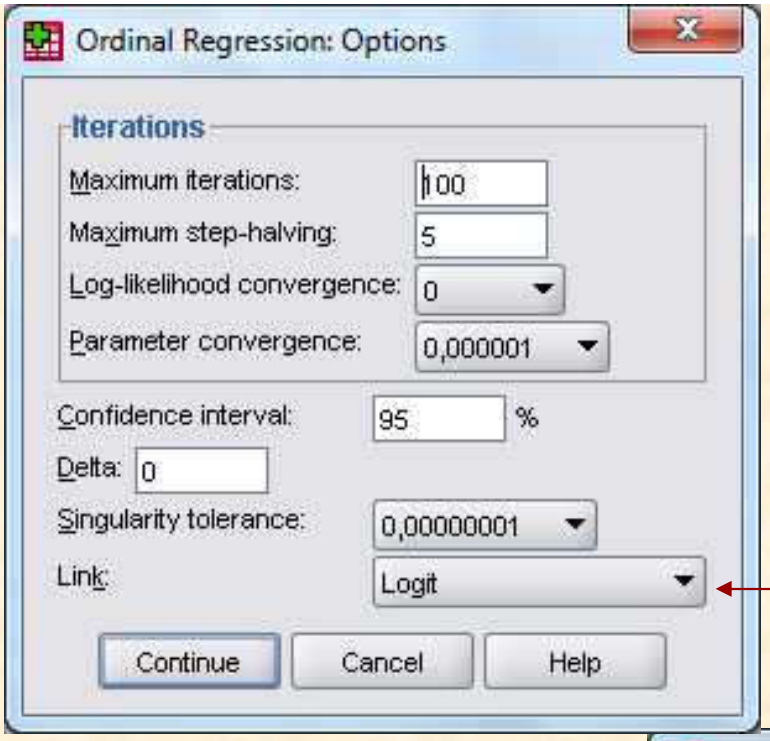
Örnek.

336 hasta üzerinde yapılan bir çalışmadan aşağıdaki sonuçlar elde edilmiştir. Hastaların risk durumları (Düşük, Orta, Yüksek) üzerinde kollestrol, tansiyon ve skor puanlarından hangilerinin önemli etkisi vardır?

Risk	Kollestrol	Tansiyon	skor
Yüksek	Yok	Yok	3,26
Orta	Var	Yok	3,21
Düşük	Var	Var	3,94
Orta	Yok	Yok	2,81
Orta	Yok	Yok	2,53
Düşük	Yok	Var	2,59
Orta	Yok	Yok	2,56
Orta	Yok	Yok	2,73
Düşük	Yok	Yok	3,00
Orta	Var	Yok	3,50
Düşük	Var	Var	3,65
Orta	Yok	Yok	2,84
Yüksek	Yok	Var	3,90
Orta	Yok	Yok	2,68
Düşük	Var	Yok	3,57
Düşük	Yok	Yok	3,09
Düşük	Yok	Var	3,50
Düşük	Yok	Yok	2,17
Yüksek	Yok	Var	3,36
Orta	Yok	Yok	3,40
Yüksek	Yok	Yok	2,75
Orta	Var	Yok	3,20
Düşük	Yok	Yok	2,44

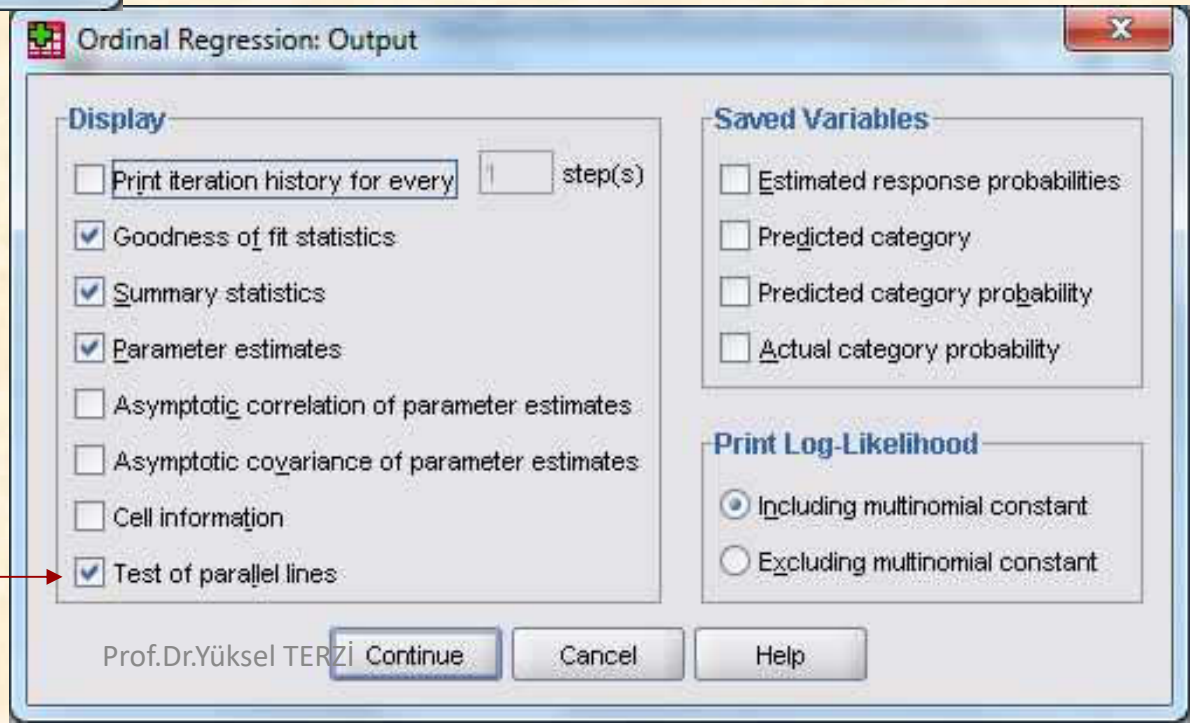

Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Linear...
Curve Estimation...
Partial Least Squares...
Binary Logistic...
Multinomial Logistic...
Ordinal...
Probit...


Ordinal Regression
Dependent:
Risk
Factor(s)
Kollestrol
Tansiyon
Covariate(s)
skor
Options...
Output...
Location...
Scale...
OK
Paste
Reset
Cancel
Help



Link fonksiyonu Logit seçilir

Paralel eğriler varsayımı için paralel seçim yapılır.



Paralel eğriler varsayımı

Test of Parallel Lines[a]

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Null Hypothesis	533,091			
General	529,077	4,014	3	,260

The null hypothesis states that the location parameters (slope coefficients) are the same across response categories.

a. Link function: Logit.

Bu sonuç, tahmin edilen regresyon katsayılarının, bağımlı değişkenin her bir kategorisinde aynı olduğunu ve **paralel eğriler varsayımının** sağlandığını (p=0,260>0,05) göstermektedir.

Modelin uygunluk testi

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	400,843	435	,878
Deviance	400,749	435	,879

Link function: Logit.

Logit bağlantılı sıralı lojistik regresyon modelinin uygun olduğu görülmektedir (p>0,05).

Pseudo R^2 (Sözde R^2)

Pseudo R-Square

Cox and Snell	,059
Nagelkerke	,070
McFadden	,033

Link function: Logit.

Bağımlı değişken ile açıklayıcı değişkenler arasındaki ilişkinin gücünü ölçmek ve değerlendirmek için elde edilen Pseudo (sözde) R^2 değerleri, Cox-Snell, Nagelkerke ve McFadden bulunmuştur. Burada **sözde R^2 değerleri, bağımlı değişkendeki değişkenliğin açıklayıcı değişkenler tarafından açıklanma oranını göstermektedir.** Ancak bu değer kesin sonuçlar vermemektedir.

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[Risk = 0]	1,214	,939	1,671	1	,196	-,627	3,055
	[Risk = 1]	3,310	,952	12,076	1	,001	1,443	5,177
Location	skor	,616	,263	5,499	1	,019	,101	1,130
	[Kollestrol=0]	-1,048	,268	15,231	1	,000	-1,574	-,522
	[Kollestrol=1]	0[a]	.	.	0	.	.	.
	[Tansiyon=0]	,059	,289	,041	1	,839	-,507	,624
	[Tansiyon=1]	0[a]	.	.	0	.	.	.

Link function: Logit.

a. This parameter is set to zero because it is redundant.

Threshold (Eşik) değerleri bağımlı değişkenin farklı kategorilerinin olasılık değerlerinin hesaplanmasında kullanılmaktadır. Risk=1 (orta) eşik değeri önemli bulunmuştur. Ayrıca açıklayıcı değişkenlerden skor puanları ve kollestrol=0 durumu risk üzerinde istatistiksel olarak anlamlı bulunmuştur.

Parametrelerin yorumu:

Tahmin edilen parametre değeri istatistiksel olarak anlamlı ve pozitif işaretli olan değişkenlerin değeri bir birim arttırıldığında, her bir değişken, bağımlı değişkende sahip olduğu parametre değeri kadar artışta neden olur.

Parametre değeri istatistiksel olarak anlamlı ve negatif işaretli olan bağımsız değişkenlerin değeri bir birim arttırıldığında, her bir değişken bağımlı değişken düzeyinde sahip olduğu parametre değeri kadar azalışa neden olacaktır.

3. İsimsel Lojistik Regresyon Analizi (Multinominal Logistic Regression)

Cevap değişkeni isimsel ölçekli olup, en az üç kategoriden oluşmalıdır (fen-tıp-eğitim-iibf gibi).

Parametre tahminleri yinelemeli-ağırlıklı en küçük kareler yöntemine (iterative-reweighted least square method) göre en büyük benzerlik parametre tahminleri yapar. Kategoriler birbirine paraleldir varsayımı kullanılır.

En büyük değere sahip cevap referans alınarak bu referansa göre lojit modeller türetilerek analiz yapılır. Referans değer belirtilmemiş ise ilk cevap referans olarak alınır. Multinomial logistic regresyon çok sayıda grubu ikili lojistik regresyonun kombinasyonu biçiminde karşılaştırır.

Multinomial logistic regresyon her bir ikili karşılaştırma için katsayılar seti sunar.Referans grup için tüm katsayılar sıfırdır.

Multinomial logistic regresyonda normallik, lineerlik ve bağımsız değişkenlerin varyanslarının homojenliği varsayımları yoktur.

Bu varsayımların gerekliliği olmadığından Multinomial logistic regresyon analizi diskriminant analizine tercih edilir.

$$pr(y_i = 1|\mathbf{x}_i) = \frac{1}{1+\sum_{j=2}^{J}\exp(\mathbf{x}_i\beta_j)} \quad \text{for m} = 1$$

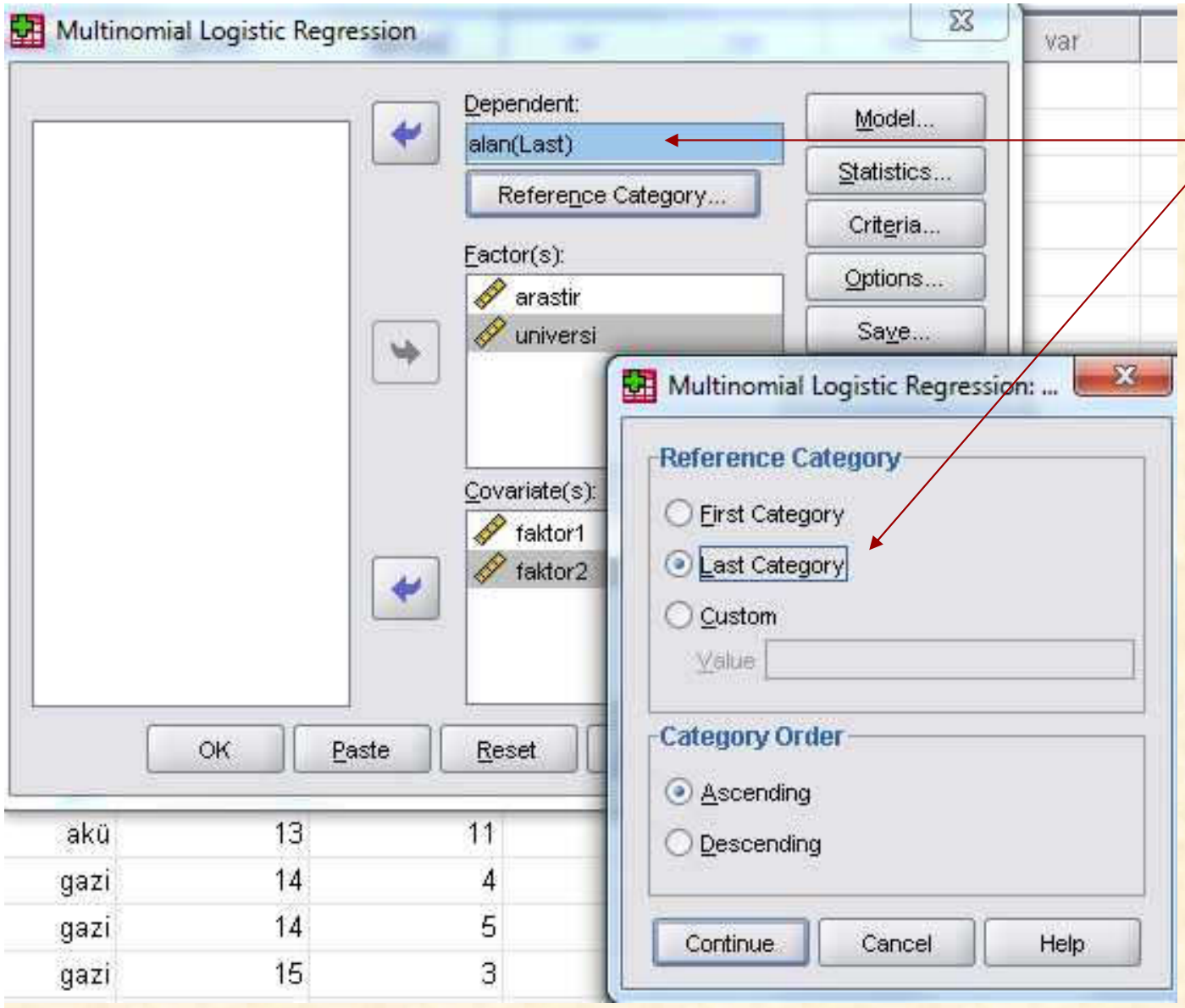
$$pr(y_i = m|\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i\beta_m)}{1+\sum_{j=2}^{J}\exp(\mathbf{x}_i\beta_j)} \quad \text{for m} > 1$$

Örnek.

196 kişi üzerinde yapılan bir araştırmada kişilerin 3 farklı alana (sosyal, fen, sağlık) göre araştırma yapıp yapmaması, bağlı oldukları üniversiteler ve yayınlardan almış oldukları faktör puanları verilmiştir. Alan üzerinde araştırma, ünv., faktör1 ve faktör2 açıklayıcı değişkenlerinden hangileri istatistiksel olarak önemlidir?

Alan:sosyal, fen,sağlık
Araştırma: Evet, hayır
Üniversite:Omü, Gazi, Akü, Ktü
Faktör1: Test puanı
Faktör2: Skor puanı

alan	arastir	universi	faktor1	faktor2
sosyal	evet	gazi	12	2
sosyal	hayir	omü	8	3
sosyal	evet	akü	10	3
sosyal	hayir	ktü	12	3
sosyal	evet	gazi	15	3
sosyal	evet	gazi	19	3
fen	hayir	akü	9	3
fen	hayir	akü	11	3
fen	hayir	akü	12	3
fen	hayir	omü	21	3
sağlık	hayir	gazi	10	3
sosyal	evet	gazi	9	4
sosyal	evet	omü	10	4
sosyal	evet	gazi	11	4
sosyal	evet	gazi	11	4
sosyal	evet	gazi	14	4
sosyal	evet	gazi	15	4
sosyal	evet	gazi	16	4
sosyal	evet	gazi	21	4
sosyal	evet	omü	23	4
fen	evet	gazi	11	4
fen	hayir	omü	15	4
fen	hayir	akü	16	4
fen	evet	gazi	19	4
sağlık	evet	gazi	12	4



Bağımlı değişkenin son şıkkı (sağlık) referans grubu alınarak diğer iki alan (sosyal ve fen) için analiz yapılabilir.

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	697,378			
Final	631,161	66,216	12	,000

Modelin uygunluğunun ki-kare testi ile test edilmesi:

H_0: Bağımsız değişkenlerin olmadığı model ile bağımsız değişkenlerin yer aldığı model arasında fark yoktur.

P=0,00<0,05 olup H_0 red edilir. Bağımsız değişkenli model bağımsız değişkensiz modelden farklıdır.

Case Processing Summary

		N	Marginal Percentage
alan	sosyal	125	35,7%
	fen	114	32,6%
	sağlık	111	31,7%
arastir	evet	184	52,6%
	hayir	166	47,4%
universi	ankara	94	26,9%
	gazi	98	28,0%
	akü	81	23,1%
	ktü	77	22,0%
Valid		350	100,0%
Missing		0	
Total		350	
Subpopulation		276[a]	

a. The dependent variable has only one value

Bağımlı değişkenin herbir bir kategorisinin oranlarının kareleri toplamı alınır.

$0.357^2 + 0.326^2 + 0.317^2 = 0.333$

Bu oran 1.25 ile çarpılır.

1.25 x 0,333 = %41.6

Classification

Observed	Predicted			
	sosyal	fen	sağlık	Percent Correct
sosyal	64	29	32	51,2%
fen	30	74	10	64,9%
sağlık	49	30	32	28,8%
Overall Percentage	40,9%	38,0%	21,1%	48,6%

Doğru sınıflama oranı (%48.6) %41.6'ya eşit yada büyük olmalıdır.

Bu örnek için doğru sınıflama kriteri yeterlidir.

Likelihood Ratio Tests

Effect	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC of Reduced Model	BIC of Reduced Model	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	659,161	713,172	631,161[a]	,000	0	.
faktor1	660,991	707,286	636,991	5,830	2	,054
faktor2	655,191	701,486	631,191	,029	2	,985
arastir	681,745	728,040	657,745	26,584	2	,000
universi	664,297	695,160	648,297	17,135	6	,009

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

Faktör1, Araştırma ve üniversite açıklayıcı değişkenleri ile bağımlı değişken (Alan) arasındaki ilişki istatistiksel olarak önemlidir (p<0,05)

Parameter Estimates

alan[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
sosyal	Intercept	,079	,738	,012	1	,914			
	faktor1	-,028	,036	,601	1	,438	,972	,906	1,044
	faktor2	,010	,057	,028	1	,866	1,010	,902	1,130
	[arastir=1]	,523	,287	3,311	1	,069	1,686	,960	2,960
	[arastir=2]	0[b]	.	.	0	.	.	.	.
	[universi=1]	,468	,405	1,339	1	,247	1,597	,723	3,531
	[universi=2]	,517	,378	1,870	1	,171	1,677	,799	3,520
	[universi=3]	-,302	,388	,606	1	,436	,739	,345	1,583
	[universi=4]	0[b]	.	.	0	.	.	.	.
fen	Intercept	1,647	,745	4,886	1	,027			
	faktor1	-,091	,039	5,484	1	,019	,913	,846	,985
	faktor2	,007	,061	,012	1	,914	1,007	,894	1,134
	[arastir=1]	-,956	,300	10,133	1	,001	,385	,214	,693
	[arastir=2]	0[b]	.	.	0	.	.	.	.
	[universi=1]	1,232	,432	8,142	1	,004	3,429	1,471	7,996
	[universi=2]	,327	,441	,548	1	,459	1,386	,584	3,291
	[universi=3]	,201	,429	,220	1	,639	1,223	,527	2,836
	[universi=4]	0[b]	.	.	0	.	.	.	.

a. The reference category is: sağlık.

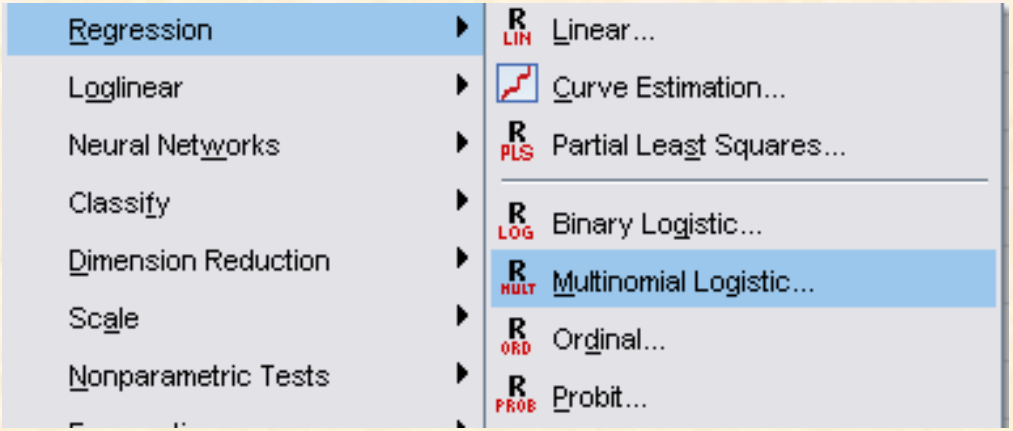
b. This parameter is set to zero because it is redundant.

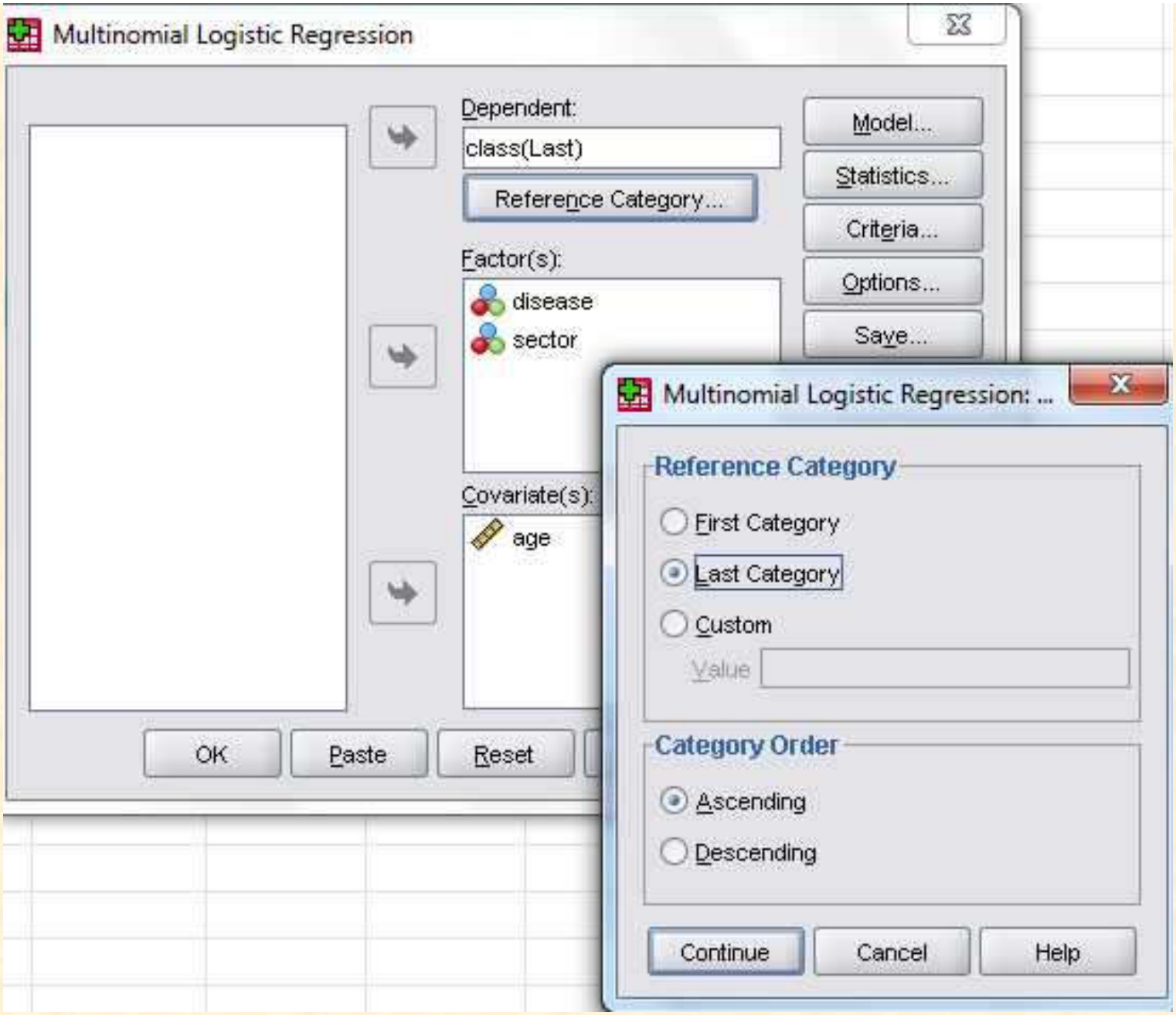
Fen alanını tercih etmede faktör1, araştırma1(evet) ve ünv=1(omü) istatistiksel olarak önemli bulunmuştur.

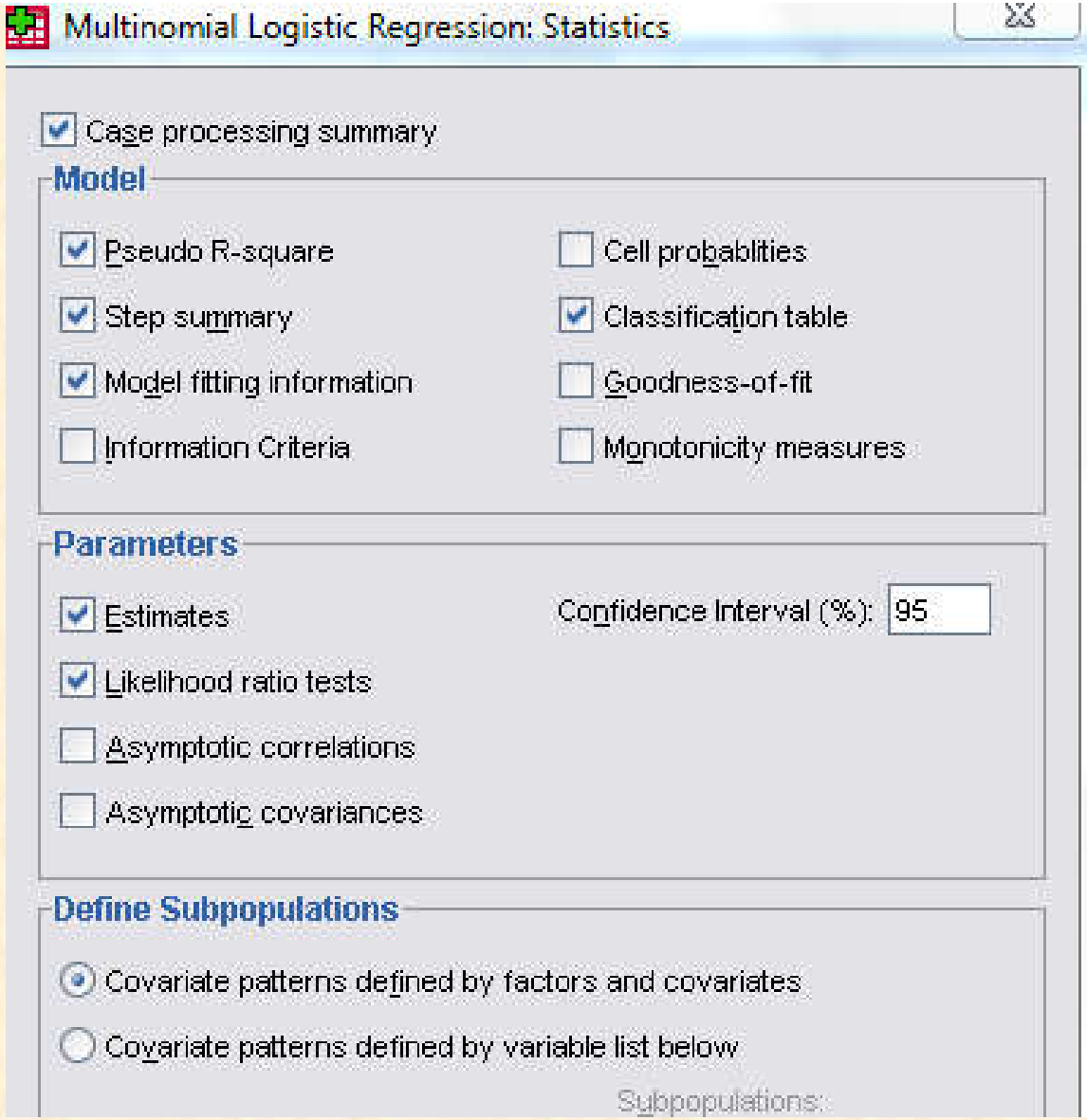
Örnek.

class	age	sector	disease
A	33	1	0
A	35	1	0
A	6	1	0
A	60	1	0
C	18	1	YOK
C	26	1	0
C	6	1	0
B	31	1	YOK
B	26	1	YOK
B	37	1	0
A	23	1	0
A	23	1	0
A	27	1	0
A	9	1	YOK
A	37	2	YOK
A	22	2	YOK
A	67	2	YOK
A	8	2	0
A	6	2	YOK
A	15	2	YOK
B	21	2	YOK
B	32	2	YOK
A	16	2	YOK
B	11	2	0

Class: A,B,C
Age: Yaş
Sector: 1,2
Disease: Yok (0), Var (1)




Multinomial Logistic Regression
Dependent:
class(Last)
Reference Category...
Factor(s):
disease
sector
Covariate(s):
age
Model...
Statistics...
Criteria...
Options...
Save...
OK
Paste
Reset
Multinomial Logistic Regression: ...
Reference Category
First Category
Last Category
Custom
Value
Category Order
Ascending
Descending
Continue
Cancel
Help


Multinomial Logistic Regression: Statistics
Case processing summary
Model
Pseudo R-square
Cell probablities
Step summary
Classification table
Model fitting information
Goodness-of-fit
Information Criteria
Monotonicity measures
Parameters
Estimates
Confidence Interval (%): 95
Likelihood ratio tests
Asymptotic correlations
Asymptotic covariances
Define Subpopulations
Covariate patterns defined by factors and covariates
Covariate patterns defined by variable list below
Subpopulations:

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	335,662			
Final	321,097	14,566	6	,024

Modelin uygunluğunun ki-kare testi ile test edilmesi:

H0: Bağımsız değişkenlerin olmadığı model ile bağımsız değişkenlerin yer aldığı model arasında ilişki yoktur.

P=0,00<0,05 olup H0 red edilir. Bağımsız değişken ile bağımlı değişken arasındaki ilişki önemlidir.

Case Processing Summary

		N	Marginal Percentage
class	A	77	39,3%
	B	49	25,0%
	C	70	35,7%
sector	1	117	59,7%
	2	79	40,3%
disease	0	139	70,9%
	YOK	57	29,1%
Valid		196	100,0%
Missing		0	
Total		196	
Subpopulation		127[a]	

Bağımlı değişkenin herbir bir kategorisinin oranlarının kareleri toplamı alınır.

$0.393^2 + 0.25^2 + 0.357^2 = 0.344$

Bu oran 1.25 ile çarpılır.

1.25 x 0,344 = %43

Classification

Observed	Predicted			
	A	B	C	Percent Correct
A	41	0	36	53,2%
B	23	0	26	,0%
C	21	0	49	70,0%
Overall Percentage	43,4%	,0%	56,6%	45,9%

Doğru sınıflama oranı (%45.9) %43'e eşit yada büyük olmalıdır.

Bu örnek için doğru sınıflama kriteri yeterlidir.

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	3,211E2	,000	0	.
age	323,115	2,019	2	,364
sector	333,135	12,039	2	,002
disease	321,514	,418	2	,811

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

Sector açıklayıcı değişkeni ile bağımlı değişken (Alan) arasındaki ilişki istatistiksel olarak önemlidir ($p<0,05$)

Parameter Estimates

class[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
A	Intercept	,428	,473	,818	1	,366			
	age	,009	,009	1,012	1	,315	1,009	,991	1,028
	[sector=1]	-1,194	,377	10,005	1	,002	,303	,145	,635
	[sector=2]	0[b]	.	.	0	.	.	.	.
	[disease=0]	,254	,404	,395	1	,530	1,289	,584	2,848
	[disease=1]	0[b]	.	.	0	.	.	.	.
B	Intercept	,291	,529	,302	1	,582			
	age	-,004	,011	,125	1	,724	,996	,975	1,018
	[sector=1]	-1,080	,417	6,710	1	,010	,339	,150	,769
	[sector=2]	0[b]	.	.	0	.	.	.	.
	[disease=0]	,201	,454	,197	1	,658	1,223	,502	2,981
	[disease=1]	0[b]	.	.	0	.	.	.	.

a. The reference category is: C.

b. This parameter is set to zero because it is redundant.

Sector1 açıklayıcı değişkeni hem A sınıflamasında hem de B sınıflamasında istatistiksel olarak önemli bulunmuştur (p<0,05). Sector1 değişkeninin parametre katsayısı negatif olduğundan exp(B) 1'den küçük çıkmıştır. A sınıfı için exp(B)=0,303 şu şekilde yorumlanır: 1-0,303=0,697 yani sector1'in, sector 2'ye göre A sınıfını tercih etmesi %69,7 daha düşüktür.