



İSTATİSTİK

8. Lojistik Regresyon

Lojistik Regresyon

- Kategorik olan bir bağımlı değişkeni bir yada daha fazla bağımsız değişken kullanarak tahminlemek istediğimizde kullanıyoruz. Bağımsız değişkenler kategorik veya sürekli olabilir.
- Bu amaçla normal regresyon kullanamıyoruz çünkü çıktı değişkeni kategorik olduğundan, normal regresyonun doğrusallık şartı yerine gelmemiş olur
- Lojistik regresyonda, bağımlı değişkenin alabileceği değerlerden birinin gerçekleşme olasılığı tahmin edilir

Varsayımlar

- ▶ Bağımlı değişken kategoriktir
- ▶ Logit denklemi doğrusaldır
- ▶ Mükemmel ayrılmanın olmaması gerekir
- ▶ Bağımsız değişkenler arasında yüksek korelasyon olmamalı (multicollinearity)
- ▶ Uç değerlerin etkisine dikkat edilmelidir
- ▶ Örneklem, yeterli büyüklükte olmalıdır

$$P(Y) = \frac{e^{b_0 + b_1 X_1 + \dots + b_n X_n}}{1 + e^{b_0 + b_1 X_1 + \dots + b_n X_n}} = \frac{e^L}{1 + e^L}$$

- ▶ Çıktı
 - ▶ Çıktının olma olasılığını hesaplıyoruz
 - ▶ Bağımsız değişkenlerin, 1 olarak kodlanan değer olma olasılığı üzerindeki etkilerini bulmaya çalışıyoruz.
- ▶ Katsayılar
 - ▶ $b_0 + b_1 X_1 + \dots + b_n X_n$ ifadesi logit olarak adlandırılır.
 - ▶ Normal regresyondaki katsayılar gibi düşünülebilir.
 - ▶ Normal regresyon denklemi lojistik regresyon denkleminin bir bölümünü oluşturuyor.

Modeli Değerlendirme

$$\log - \text{likelihood} = \sum_{i=1}^N [Y_i \ln(P(Y_i)) + (1 - Y_i) \ln(1 - P(Y_i))]$$

- The Log-likelihood istatistiği
 - Çoklu regresyondaki hatalar kareleri toplamına benzer
 - Model veriye uygulandıktan sonra ne kadarlık bir açıklanmayan varyans olduğunu gösterir
 - Eğer elde edilen değer büyükse modelin iyi uyum sağlamadığı anlaşılır.
 - Lojistik regresyonda, çoklu regresyon R^2 sine benzer bazı R^2 değerleri hesaplanır. Ancak bunların yorumlanmasına dikkat etmek gerekir.

Farklı Modelleri Değerlendirme

- Farklı modeller için ayrı ayrı log-likelihood istatistikleri hesaplamak ve bu log-likelihood değerleri arasındaki farka bakarak modelleri karşılaştırmak mümkündür.

$$\chi^2 = 2[LL(New) - LL(Baseline)]$$

$$(df = k_{new} - k_{baseline})$$

Modeli Değerlendirme

- ▶ Hosmer ve Lemeshow testinin sonucu anlamsız çıkarsa (bu durumda gözlemlenen ve model tarafından kestirilen değerler arasında anlamlı bir farklılık yoktur), lojistik regresyon modelinin bir bütün olarak veriye yeterince uyum sağladığı sonucuna varılır.
- ▶ Ayrıca modelin doğru tahminleme gücü de ne kadar iyi bir model olduğunu anlamak için kullanılır.

Bağımsız Değişken Katsayılarını Değerlendirme: The Wald Statistic

- ▶ Regresyon denklemindeki t -istatistiğine benzer
- ▶ $b = 0$ hipotezini test eder.
- ▶ b büyük olduğunda ürettiği sonuçlara güvenilmeyebilir.

$$Wald = \frac{b}{SE_b}$$

Regresyon Katsayılarının Anlamı

$$\text{Exp}(b) = \frac{\text{Bir birimdeğişikliksonrasiodds}}{\text{Değişimöncesiiodds}}$$

- ▶ Bağımsız değişkende bir birim değişikliğe bağlı olarak odds'ta görülen değişikliği ifade eder
 - ▶ b pozitifse veya $\text{Exp}(b) > 1$: Bağımsız değişken $\uparrow$, Çıktının olma olasılığı $\uparrow$.
 - ▶ b negatifse veya $\text{Exp}(b) < 1$: Bağımsız değişken $\uparrow$, Çıktının olma olasılığı $\downarrow$.
- ▶ Bağımsız değişkendeki bir birimlik değişime karşılık bağımlı değişkenin gerçekleşme odds'unda $\% (\text{Exp}(b)-1)*100$ değişme olur.