İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi Yıl:4 Sayı:7 Bahar 2005/1 s. 59-72

POISSON REGRESYON ANALİZİ

Özlem DENİZ*

ÖZET

Herhangi bir olayın belirlenen bir süreç içerisinde yapılan denemeler sonucunda meydana gelme sayısı, sayma verileri olarak ifade edilebilir. Sayma veri modelinde bilinen ilk gelişmeler aktueryal bilimler, biyoistatistik ve demografide gözlenmiştir. Son yıllarda bu modeller iktisat, politik bilimler ve sosyolojide de sıkça kullanılmaya başlanmıştır. Sayma veri modelleri özel bir regresyon türüdür. Bu modeller ekonometricilerin çok fazla dikkatini çekmiş ve mikro ekonomide oldukça fazla kullanılmıştır. Bilindiği gibi, verilerin sürekli olduğu durumlarda doğrusal regresyon analizi kullanılabilmektedir. Ancak analizlerde kullanılacak veriler her zaman sürekli halde bulunmayabilir. Bu gibi durumlarda yani; verilerin kesikli olması durumunda da doğrusal regresyon modelleri kullanılarak yapılacak analizler etkisiz, tutarsız ve çelişkili sonuçlar verecektir. Bu sebepten dolayı kesikli veriler için tüm koşullar sağlandığında kulanılabilecek en etkin model Poisson regresyon modelleridir.

Anahtar Kelimeler: Poisson Regresyon, Yapay En Çok Olabilirlik Kestirimi, Artık Analizi

POISSON REGRESSION ANALYSIS

ABSTRACT

The occurance number (frequency) of an event tested in a determined progress is called counting data. The first improvements in counting data model were seen in actuarial sciences, biostatistics and demography. Counting data models are a specific kind of regression. As we all know, linear regression can be used where the data is continuous. However the data can not always be continuous. In these circumstances where the data is discontinuous, the application of linear regression leads us to ineffective, incosistent and contradictory results. Therefore, when all the conditions for discontinuous data are met, Poisson regression models are the most effective model.

Keywords: Poisson Regression, Artificial Maximum Likelihood Prediction, Residual Analysis

* *İstanbul Ticaret Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü, odeniz@iticu.edu.tr*

1. GİRİŞ

Herhangi bir olayın belirlenen bir süreç içerisinde yapılan denemeler sonucunda meydana gelme sayısı, sayma verileri olarak ifade edilebilir. Sayma veri modelinde bilinen ilk gelişmeler aktueryal bilimler, biyoistatistik ve demografide gözlenmiştir. Son yıllarda bu modeller iktisat, politik bilimler ve sosyolojide de sıkça kullanılmaya başlanmıştır.

Sayma veri modelleri özel bir regresyon türüdür. Bu modeller ekonometricilerin çok fazla dikkatini çekmiş ve mikro ekonomide oldukça fazla kullanılmıştır.

Sayma verilerinin analizi için ilk sorulan soru "özel" yöntemlerin gerekliliği veya doğrusal regresyon modelinin yeterli olup olmadığıdır. Sayma verilerinden oluşan değişkenler için sürekli ve doğrusal regresyon modelinin uygulanabileceği düşünülür. Ancak bu verilere doğrusal regresyon modeli uygulanması halinde sonuçlar, etkisiz ve tutarsız olduğu gibi çelişkili tahminleri yapılabilir. Sayma sonuçlarının özelliklerini kesin olarak veren birçok model vardır. Ancak Poisson regresyon birçok analizin başlangıç noktası olarak düşünülür. Poisson regresyon modeli sayma verileri için en sık kullanılan ve en basit olan yöntemdir. Bu model ile sayımın olasılığı, Poisson dağılımı ile belirlenir. Bu modelin belirgin özelliği, sonucun koşullu ortalamasının koşullu varyansına eşit olmasıdır. Ancak uygulamada bazen koşullu varyans, koşullu ortalama değerini aşabilir. İşte bu tür durumlarda, negatif binom regresyon modelleri kullanılır.

Bu çalışmada, koşullu ortalamanın koşullu varyansa eşit olduğu durumda kullanılan Poisson regresyon analizi, teorik olarak açıklanmaya çalışılmıştır.

2. POISSON REGRESYON SÜRECİ

Bağımlı değişkenin 0, 1, 2, 3, ... gibi kesikli değer aldığı fakat kategorik olmadığı durumlar vardır. Bu tür değişkenlere, doğalgaz boruları üzerinde kazaların sayısı, verilen patentlerin sayısı, yazlıklarda çıkan yangınların sayısı gibi örnekler gösterilebilir. Kesikli ve kategorik olmayan, nadir olaylarla ilişkili bağımlı değişkenli model, bazı varsayımlar altında Poisson regresyon modeli olarak adlandırılır. Poisson regresyon modeli daha çok sayma verilerini analiz etmek için kullanılmaktadır (Akın, 2002).

Poisson regresyon modelinde regresyon sürecindeki genel kestirimler en çok olabilirlik yöntemi ile gerçekleştirilmektedir. Poisson en çok olabilirlik kestirimi için;

1) Koşullu ortalamanın doğru tanımlanmasında bağımlılık şartı sağlanmalıdır. Ayrıca bağımlı değişken y 'nin Poisson dağılması gereklidir.
2) En çok olabilirlik standart hataları ve t istatistikleri kullanarak hesaplanan istatistiksel sonuçlar, hem koşullu ortalama, hem varyansın doğru tanımlanmasını gerektirmektedir. Burada istenen koşul, koşullu varyans ve ortalamanın eşit olmasıdır.
3) Veriler için koşullu varyans ve koşullu ortalamanın eşit olmaması durumunda, en çok olabilirlik yönteminin uygulanması ile elde edilmiş istatistiksel sonuçlar, koşullu ortalamanın doğru tanımlandığının ispat edildiği durumlarda geçerli ve doğrudur.
4) Veriler için koşullu varyans ve ortalamanın eşit olmaması durumunda, Poisson en çok olabilirlik tahmin edicisinden daha etkin tahmin ediciler kullanılabilir.

2.1. Poisson Regresyon Sürecinde Katsayıların Kestirimi

Poisson regresyon sürecinde bağımlı değişken y_i 'nin dağılımına göre, $\hat{\beta}$ kestiricilerini hesaplama yöntemleri değişiklik göstermektedir. En çok olabilirlik kestirim yöntemi (MLE), doğrusal ve karesel varyans fonksiyonları ile negatif binom, yapay en çok olabilirlik (PMLE) ve genelleştirilmiş doğrusal modeller, bu yöntemlerden en çok bilinen ve en sık kullanılanlarıdır.

2.1.1. Poisson En Çok Olabilirlik Kestirim Yöntemi

x_i 'ye bağlı y_i için Poisson regresyon modeli;

$$f(y_i|x_i)=\frac{e^{-\mu_i}\ \mu_i^{y_i}}{y_i!}\ ,\ y_i=0,1,2,... \tag{1}$$

ve ortalama parametresi;

$$E[y_i|x_i]=\mu_i=\exp(x_i'\beta) \tag{2}$$

şeklinde gösterilir ve "üstel ortalama fonksiyonu" olarak ifade edilir. İstatistik literatüründe bu fonksiyon ayrıca; "log-doğrusal fonksiyon" olarak da ifade edilir. Çünkü koşullu ortalamanın logaritması, parametreleri doğrusal olarak vermektedir.

$$\ln E[y_i|x_i]=\mu_i=x_i'\beta \tag{3}$$

Bağımsız gözlemler için, log-olabilirlik fonksiyonu;

$$\ln L(\beta) = \sum_{i=1}^{n} \left\{ y_i x_i' \beta - \exp\left(x_i' \beta\right) - \ln y! \right\} \tag{4}$$

Buna bağlı olarak Poisson MLE $\hat{\beta}_p$ değeri;

$$\sum_{i=1}^{n} \left(y_i - \exp\left(x_i' \beta\right) \right) x_i = 0 \tag{5}$$

ifadesinden bulunur.

$\hat{\beta}_p$ değerinin hesaplanmasında kullanılan standart yöntem, Fisher iterasyon yöntemidir. Uygulamada genellikle 10 veya daha az iterasyon yapmak yeterli olmaktadır.

Verilen bilgiler uygulanan modeller doğrultusunda katsayıların kestirimi için;

$$\hat{\beta}_p \overset{a}{\sim} N\left[\beta, V_{ML}\left[\hat{\beta}_p\right]\right] \tag{6}$$

ve varyans değeri için;

$$V_{ML}\left[\hat{\beta}_p\right] = \left(\sum_{i=1}^{n} \mu_i x_i x_i' \right)^{-1} \tag{7}$$

sonuçlarına ulaşılır.

2.1.2. Yapay En Çok Olabilirlik Kestiricisi

Bağımlı değişken y_i 'nin Poisson dağılıma uygunluk göstermemesi durumunda bile, Poisson regresyon yardımıyla hesaplanmış $\hat{\beta}_p$ 'ler kullanılabilir. Bu amaçla "yapay en çok olabilirlik kestiricisi" olarak adlandırılan kestiriciler kullanılır. Bu terminoloji, Poisson modelindeki Poisson en çok olabilirlik kestiricisinin, birinci dereceden koşul tanımıyla elde edilmesi gereken kestirici yerine kullanılması anlamına gelir. Ama bu kestiricinin, Poisson en çok olabilirlik kestiricisindeki gibi, Poisson dağılımına uygunluk göstermesini gerektirmez.

Bu açıklamalara bağlı olarak, Poisson için yapay en çok olabilirlik kestiricisi $\hat{\beta}_p$; varyansın,

$$V_{PML}\left(\hat{\beta}_p\right)=\left(\sum_{i=1}^{n}\mu_i x_i x_i^{'}\right)^{-1}\left(\sum_{i=1}^{n}w_i x_i x_i^{'}\right)\left(\sum_{i=1}^{n}\mu_i x_i x_i^{'}\right)^{-1} \quad (8)$$

olarak ifade edildiği;

$$\hat{\beta}_p \sim N\left[\beta, V_{PML}\left(\hat{\beta}_p\right)\right] \quad (9)$$

şeklinde dağılır. ω_i değerinin, y_i için koşullu varyans değeri olduğu bilinmektedir.

2.1.3. Genelleştirilmiş Doğrusal Modeller Kestirim Yöntemi

$E\left[y_i \mid x_i\right]=\mu_i=\exp(x_i^{'}\beta)$ ortalama fonksiyonuna sahip Poisson için, bu modelin kanonik bağ fonksiyonu olan Poisson yoğunluk fonksiyonu;

$$f\left(y_i \mid x_i\right)=\exp\left\{\frac{x_i^{'}\beta y_i-\exp\left(x_i^{'}\beta\right)}{\phi}+c\left(y_i,\phi\right)\right\} \quad (10)$$

şeklinde tanımlanır. Bu modelde $c\left(y_i,\phi\right)$, normalleştirme katsayısıdır. ϕ değeri doğrusal varyans fonksiyonu ile negatif binom dağılımı yardımıyla hesaplanmış olan $V\left[y_i\right]=\phi\mu_i$ fonksiyonundan hesaplanmaktadır.

Genelleştirilmiş doğrusal modeller yardımıyla hesaplanan Poisson kestiricisi $\hat{\beta}_{GLM}$, birinci dereceden koşullar ile;

$$\sum_{i=1}^{n}\frac{1}{\phi}\left(y_i-\exp\left(x_i^{'}\beta\right)\right)x_i=0 \quad (11)$$

denkleminden hesaplanmaktadır (Cameron ve Trivedi, 1998).

2.2. Regresyon Sonuçlarının Kullanılması

Bir önceki bölümde kullanılan yöntemler yardımıyla hesaplanan katsayılar doğru bir şekilde yorumlanmadığı sürece model için hiçbir anlam ifade etmemektedir. Ayrıca hesaplanan bu değerler yardımı ile bağımlı değişken y_i değerleri için de kestirimler yapılmalıdır. Bu bölümde regresyon katsayılarının yorumlanması ve bağımlı değişkenin kestirimi konularına değinilecektir.

2.2.1. Katsayıların Yorumlanması

Regresyon katsayılarının yorumlanması, regresyon sürecindeki önemli konulardan biridir. Örneğin; $\hat{\beta}_j$'nin 0,2 olmasının ne anlama geldiğinin açıklanması gerekmektedir. Doğrusal regresyon modelinde beklenen değer; $E\left[y \mid x\right]=x^{'}\beta$ şeklinde

hesaplanmaktaydı. Bu ifadedeki β değeri yalnız bırakılır ve $\partial E[y|x]/\partial x_j = \beta_j$ işlemi gerçekleştirilirse; $\hat{\beta}_j = 0{,}2$ için, " j 'inci bağımsız değişkendeki 1 birimlik değişim, koşullu ortalamayı 0,2 birim artırmaktadır" yorumu yapılır. Ancak Poisson regresyon modeli üstel bir yapı taşıdığı için katsayıların yorumlanması bu kadar kolay olmayacaktır.

Üstel koşullu ortalama;

$$E[y|x] = \exp(x'\beta) \tag{12}$$

şeklinde gösterilmekteydi. x_j değeri için j 'inci bağımsız değişken olduğu düşünülsün. Benzer işlemlerin tekrarlanması sonucunda;

$$\frac{\partial E[y|x]}{\partial x_j} = \beta_j \exp(x'\beta) \tag{13}$$

sonucuna ulaşılır. Örneğin, eğer $\hat{\beta}_j = 0{,}2$ ve $\exp(x'\hat{\beta}) = 2{,}5$ ise; j 'inci bağımsız değişkendeki bir birimlik değişim, y bağımlı değişkeninde 0,5 birimlik artışa neden olacağı, eşitlikten hesaplanabilmektedir (McCullagh ve Nelder, 1983).

2.2.2. Kestirilmiş Değerin Hesaplanması

Gözlem değerlerinden oluşan x bağımsız değişkeni x_p, koşullu ortalamanın tahmini değeri de $\mu_p = E[y|x = x_p]$ olarak gösterilsin.

Tanımlanan ifadeler doğrultusunda üstel koşullu ortalama fonksiyonu için, ortalamanın tahmini;

$$\hat{\mu}_p = \exp(x_p'\hat{\beta}) \tag{14}$$

şeklinde hesaplanır.

Bu değer %95 güven aralığı için;

$$\mu_p \in \hat{\mu}_p \mu z_{0,25}\sqrt{\hat{\mu}_p^2 x_p' V[\hat{\beta}] x_p} \tag{15}$$

aralığında yer almaktadır. $\hat{\beta}$ kestiricisinin; $\hat{\beta} \sim N\left[\beta, V\left[\hat{\beta}\right]\right]$ olduğu bilinmektedir. Daha dar güven aralıklarında β için daha kesin tahminler yapılabilmektedir.

Bağımlı değişken y için, ortalamanın tahmini yerine gerçek değerin tahmini istenilebilir. Gözlemler $x = x_p$ olarak tanıtıldığında, üstel koşullu ortalama formülü olarak hesaplanan tahminler;

$$\hat{y}_p = \exp\left(x_p' \hat{\beta}\right) \tag{16}$$

eşitliğinden elde edilir.

Poisson model için varyans fonksiyonu dikkate alınırsa, y_p'nin kestirilen varyansı $\omega\left(\hat{\mu}_p, \hat{\alpha}\right)$ olarak ifade edilir. Bu durumda y_p için;

$$y_p \in \hat{y}_p \pm z\sqrt{\omega\left(\hat{\mu}_p, \hat{\alpha}\right) + \hat{\mu}_p^2 x_p' V\left[\hat{\beta}\right] x_p} \tag{17}$$

aralığında olduğu söylenebilir (a.g.e., Cameron ve Triverdi, 1998).

2.3. Artıkların Analizi

Artıklar, bağımlı değişken için gerçek değerler ile kestirilmiş değerler arasındaki farka eşittir. Artıklar uç değerleri belirlemede, zayıf uyum gösteren gözlemleri kestirebilmekte, etkin gözlemleri tespit etmede ve etkin gözlemleri seçebilmede kullanılabilirler.

Doğrusal modellerde artıklar, gerçek ve kestirilen değerler arasındaki fark olarak ifade edilmektedir. Ancak doğrusal olmayan modeller için artık tanımı bir tane değildir. Poisson ve diğer genelleştirilmiş doğrusal modeller için artıklar farklı yollarla ve farklı adlarla hesaplanır.

Genel anlamda artıklar

$$r_i = \left(y_i - \hat{\mu}_i\right) \tag{18}$$

olarak ifade edilir. Burada uyum ortalaması $\hat{\mu}_i = \mu\left(x_i' \beta\right)$'nın koşullu ortalamasıdır.

Normal dağılımlı klasik doğrusal regresyon modelinde homoskedastik hata $\left(y - \mu\right) \sim N\left[0, \sigma^2\right]$ olarak tanımlanır. Böylece geniş örneklerde artıklar sabit varyans ile "0" etrafında simetrik olarak dağılırlar. Sayma verileri için ise $\left(y - \mu\right)$, heteroskedastik ve asimetriktir. Böylece geniş örnekler için hata terimleri heteroskedastik ve asimetrik olduğu söylenebilir.

Sayma verileri için sıfır ortalama, sabit varyans ve simetrik dağılıma sahip bir artık yoktur.

Yapılan düzenlemeler sonucunda heteroskedasite probleminden kurtarılmış artıklar "Pearson artıklar" olarak adlandırılır ve

$$P_i = \frac{(y_i - \hat{\mu}_i)}{\sqrt{\hat{\omega}_i}} \tag{19}$$

şeklinde hesaplanır. $\hat{\omega}_i$; bağımlı değişkenin ω_i varyansının kestirimidir. Bu artıkların kareleri toplamı Pearson istatistiklerinde kullanılır. Poisson modellerde $\omega = \mu$, genelleştirilmiş doğrusal modellerde $\omega = \alpha\mu$ ve karesel varyans fonksiyonuna sahip negatif binom modellerinde $\omega = \mu + \alpha\mu^2$ olarak hesaplanır. Pearson artık değerleri "0" ortalama ve homoskedasiteye sahiptir. Ancak bu değerlerin asimetrik dağılıma sahip olduğu belirtilmelidir.

Eğer y, doğrusal üstel aile yoğunluk fonksiyonu olarak hesaplanırsa, "sapma artıklar" kullanılır ve

$$d_i = \text{sign}(y_i - \hat{\mu}_i)\sqrt{2\{\lambda(y_i) - \lambda(\hat{\mu}_i)\}} \tag{20}$$

şeklinde ifade edilir. $\lambda(\hat{\mu}_i)$; $\mu = \hat{\mu}$ olarak ifade edildiğinde y için belirlenmiş logaritmik yoğunluk fonksiyonu, $\lambda(y)$; $\mu = y$ olarak ifade edildiğinde y için belirlenmiş logaritmik yoğunluk fonksiyonudur. Hesaplanan bu artık değerlerinin karelerinin toplamı sapma istatistiğinde kullanılmaktadır.

Varyansı σ^2 olduğu bilinen normal dağılım altında; $d_i = (y_i - \mu_i)/\sigma$ işlemiyle standartlaştırılmış artıklara ulaşılır. Poisson için bu artıklar;

$$d_i = \text{sign}(y_i - \hat{\mu}_i)\sqrt{2\{y_i \ln(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\}} \tag{21}$$

olarak ifade edilir. Bu eşitlikte eğer $y = 0$ ise $y \ln y = 0$ olacağı görülmektedir (Long, 1997).

2.4. Uyum İyiliği

Genelleştirilmiş doğrusal modeller için en sık kullanılan uyum iyiliği ölçüleri, Pearson ve Sapma istatistikleridir. Bu ölçülerin kullanılması ile elde edilen sonuçlar, regresyon katsayılarındaki kestirim hatalarının kontrolü için, ki-kare uyum iyiliği testinde kullanılırlar.

2.4.1. Pearson İstatistiği

μ_i ortalamalı ve ω_i varyanslı bağımlı değişken y_i'ye ait herhangi bir model için standart uyum iyiliği ölçüm yöntemi pearson istatistiğidir ve

$$P = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu})^2}{\hat{\omega}_i} \tag{22}$$

olarak ifade edilir. Bu değer serinin yayılımının aşırı olup olmadığını belirlemede kullanılır. Burada $\hat{\mu}_i$ ve $\hat{\omega}_i$ değerleri, μ_i ve ω_i'nin kestirim değerleridir. Hesaplanan P değeri, $\hat{\mu}$ için belirlenmiş serbestlik derecesi $(n-k)$ ile karşılaştırılır.

Bu formül Poisson regresyon için uygulandığında, $\omega_i = \mu_i$ olacaktır ve

$$P_p = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \tag{23}$$

şeklini alacaktır. Hesaplanan P_p değeri de benzer şekilde $(n-k)$ değeri ile karşılaştırılacaktır. Burada;

$P_p > n-k \quad \Rightarrow$ seride aşırı yayılım

$P_p < n-k \quad \Rightarrow$ seride eksik yayılım

olduğu söylenir.

2.4.2. Sapma İstatistiği

Uyum iyiliğinin ölçülmesinde kullanılan diğer bir teknik de sapma istatistiğidir. Bu istatistik değerine aynı zamanda "G kare istatistiği" de denilmektedir.

G kare istatistiği;

$$G^2 = 2\sum_{i=1}^{n} y_i \ln\left(\frac{y_i}{\mu_i}\right) \tag{24}$$

şeklinde ifade edilir. Bu istatistik değeri 0'a yakınsıyor ise model uyumu artıyor denilebilir. Eğer bu istatistik değeri tam 0'a eşit ise model uyumunun mükemmel olduğu söylenebilir.

2.4.3. Yapay R^2 Ölçümü

Doğrusal olmayan modeller için kullanılan ortak bir R^2 tanımı bulunmamaktadır. Bu belirsizlik yüzünden hesaplanan değer ifade edilirken "yapay" ifadesi kullanılmaktadır.

Doğrusal regresyon modellerinde, R^2'nin hesaplanması için başlangıç noktası genel kareler toplamlarının ayrıştırılmasıdır. Genel olarak;

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(y_i - \hat{\mu}_i\right)^2 + \sum_{i=1}^{n}\left(\hat{\mu}_i - \bar{y}\right)^2 + 2\sum_{i=1}^{n}\left(y_i - \hat{\mu}_i\right)\left(\hat{\mu}_i - \bar{y}\right) \tag{25}$$

ifadesinde, ilk ifade genel kareler toplamı (TSS), ikinci ifade artık kareler toplamı (RSS) ve üçüncü ifade açıklanmış kareler toplamı (ESS) olarak açıklanır. Son ifade ise eğer model sabit terim içeriyorsa, doğrusal regresyon modelinin en küçük kareler kestirimine göre sıfıra eşit olacaktır. Ancak Poisson'u da içeren ve doğrusal olmayan en küçük kareler ile üstel koşullu ortalamaya sahip tüm kestiriciler ve modeller için sıfıra eşit olmayacaktır. Bu durum da R^2'nin, $R^2 = 1 - RSS/TSS$ veya $R^2 = ESS/TSS$ yönteminden farklı bir yolla hesaplanması gerektiğini ortaya çıkarmıştır (Cameron ve Triverdi, 1998).

Normallik varsayımı gerektirmeyen Poisson regresyon modeline R^2 ölçüsü olabilirlik oran yaklaşımına dayanmaktadır. Doğrusal regresyon modeline ilişkin EKK tahmini, artık kareler toplamının en çok olabilirlik tahmini ve sapma değeri ile benzer özellikler göstermesi nedeniyle önerilen R_p^2 ölçüsü;

$$R_p^2 = 1 - \frac{\log L(y) - \log L(\hat{\mu})}{\log L(y) - \log L(\bar{y})} \tag{26}$$

şeklinde tanımlanmaktadır. Burada $\log L(y)$, doygun modelin log-olabilirliğini, $\log L(\hat{\mu})$, ilgilenilen modelin log-olabilirliğini ve $\log L(\bar{y})$, sadece sabit terimin bulunduğu minimal modelin log-olabilirliğini göstermektedir. $y_i \geq 0$ gözlenen değerler, $\hat{\mu}_i = \exp(x_i \hat{\beta})$ ya da $\hat{\mu}_i = c_i \exp(x_i \hat{\beta})$ tahmin edilen değerler ve $\bar{y}_i = \exp(\hat{\beta}_0)$ ya da $\bar{y}_i = c_i \exp(\hat{\beta}_0)$ ortalama değerler olmak üzere log-olabilirlik fonksiyonları,

$$\log L(y) = \sum_{i=1}^{n} (y_i \log(y_i) - y_i - \log(y_i!)) \tag{27}$$

$$\log L(\hat{\mu}) = \sum_{i=1}^{n} (y_i \log(\hat{\mu}_i) - \hat{\mu}_i - \log(y_i!)) \tag{28}$$

$$\log L(\bar{y}) = \sum_{i=1}^{n} (y_i \log(\bar{y}_i) - \bar{y}_i - \log(y_i!)) \tag{29}$$

biçiminde elde edilmektedir. Bu log-olabilirlik fonksiyonları düzenlenirse *yapay* R^2 ölçüsüne ulaşılmaktadır (Özmen, 2003).

2.4.4. Ki-Kare Uyum İyiliği Testi

Verilmiş olan Poisson regresyon modeli için $y_i = 0,1,...,m$ olsun. Bu model için gözlenen frekanslar $\bar{p}_j$ ve teorik frekansları $\hat{p}_j$, $j = 0,1,...,m$ şeklinde ifade edilmiş olsun. Uygun bir test uygulanmadığı sürece $\hat{p}_j$'lerin $\bar{p}_j$'lere yakınlığının yeterli olup olmadığını, dolayısıyla kurulan modelin uygun olup olmadığına karar verilemez.

Uyum iyiliğini incelemek için kurulan hipotezler;

H_0 : Veriler Poisson modele uygunluk göstermektedir

H_1 : Veriler Poisson modele uygunluk göstermemektedir

şeklinde kurulabilir.

Pearson χ^2 test istatistiği;

$$\chi^2 = \sum_{i=1}^{j} \frac{\left(n\overline{p}_j - n\hat{p}_j\right)^2}{n\hat{p}_j} \tag{30}$$

"ki-kare uyum iyiliği testi" olarak adlandırılır. Bu formül yardımıyla bulunan sonuç $(N-p)$ serbestlik dereceli χ^2 değeriyle karşılaştırılır. N, birim sayısı, P, tahmin edilmek istenen parametre sayısıdır. Hesaplanan değer $\chi_{N-p}{}^2$ değerini aşıyorsa hipotez reddedilir ve verilen poisson modele uygunluk göstermediği kabul edilir (Dobson, 2002).

2.5. Regresyon Katsayılarının Anlamlılığının Testi

Hesaplanmış olan katsayıların $b_1, b_2, ..., b_k$ şeklinde gösterildiği varsayılsın. Hesapları bu katsayıların hiçbir işlem uygulamadan yorumlanmasının doğru olmadığı belirtilmişti. Çünkü kestirilen değerler, üstel fonksiyon yardımıyla türetilmişti.

Katsayıların anlamlılığının testi için kullanılacak hipotezler;

$H_0 : \beta_i = 0 \quad , (i = 1, ..., n)$ (β_i katsayısı anlamsızdır)

$H_0 : \beta_i \neq 0 \quad , (i = 1, ..., n)$ (β_i katsayısı anlamlıdır)

şeklindedir. Bu hipotezlerin testinde en sık kullanılan yöntem Wald'ın χ^2 istatistiğidir ve

$$\chi_w{}^2 = \left(\frac{b_i}{s'_{bi}}\right)^2 \tag{31}$$

şeklinde hesaplanır. Bu eşitlikte b_i, regresyon katsayılarını; s'_{bi} ise, basit standart hata değerinin ϕ sayısının karekökü ile çarpımı yardımıyla elde edilir.

$$s'_{bi} = s_{b1}\sqrt{\phi} \tag{32}$$

şeklinde ifade edilir. Böylece düzeltilmiş standart hata değerine ulaşılır. ϕ sayısı ise, k kestirilecek parametre sayısı olmak üzere;

$$\phi = \frac{1}{n-k}\sum_{i=1}^{n}\frac{(y_i-\mu_i)^2}{\mu_i} \tag{33}$$

eşitliğinden elde edilir.

Hesaplanan Wald'ın χ^2 istatistik değeri, 1 serbestlik dereceli χ^2 değeriyle karşılaştırılır. Eğer hesaplanan değer tablo değerini aşıyorsa H_0 hipotezi reddedilir. Yani katsayıların anlamlı olduğuna karar verilir.

Katsayıların anlamlılığının testinden sonra;

$$b_i \mu z_{1-\alpha/2} s'_{bi} \tag{34}$$

ifadesinin yardımıyla, katsayılar için alt ve üst limit değerleri hesaplanır.

3. SONUÇ

Bilindiği gibi, verilerin sürekli olduğu durumlarda doğrusal regresyon analizi kullanılabilmektedir. Ancak analizlerde kullanılacak veriler her zaman sürekli halde bulunmayabilir. Bu gibi durumlarda yani; verilerin kesikli olması durumunda da doğrusal regresyon modelleri kullanılarak yapılacak analizler etkisiz, tutarsız ve çelişkili sonuçlar verecektir. Bu sebepten dolayı kesikli veriler için tüm koşullar sağlandığında kulanılabilecek en etkin model Poisson regresyon modelleridir. Bu modellerin kullanılabilmesi için dikkat edilmesi gereken en önemli koşul, koşullu varyans değerinin koşullu ortalama değerine eşit olmasıdır.

Bir çok uygulamada koşullu varyans değeri, koşullu ortalama değerini aşar. Böyle durumlarda Poisson regresyonun kullanılması doğru değildir. Bunun yerine negatif binom regresyon kullanılır. Negatif binom dağılımında varyansın, ortalamanın karesel fonksiyonu olduğu varsayılır.

Poisson regresyon modeli üstel bir model olması sebebiyle katsayı yorumlamalarında zorluk ve karmaşıklık yaratması dezavantajının yanında, bağımlı değişkenin sayma verilerinden oluştuğu durumlarda doğrusal regresyon analizine alternatif olabilen bir modeldir. Bu sebeple son yıllarda pek çok alanda kullanım imkanı bulabilmektedir.

KAYNAKÇA

Akın, F., (2002), Kalitatif Tercih Modelleri Analizi, Bursa, Ekin Kitabevi.

Cameron, C.- Trivedi, P., (1998), Regression Analysis of Count Data, Cambridge, Cambridge University Pres.

Dobson, A., (2002), An Introduction to Generalized Linear Models, Boca Raton, Chapman and Hall.

Long, S., (1997), Regression Models for Categorical and Dependent Variables, London, Sage Publications.

McCullagh, P.- Nelder, J.A., (1983), Generalized Linear Models, London Chapman and Hall.

Özmen, İ., (2003), Poisson Regresyon Modeli için Düzeltilmiş Belirtme Katsayıları, Antalya İstatistik Sempozyumu Bildirisi.