S.Ü. FEN FAKÜLTESİ FEN DERGİSİ

Ekim(2020) 46(2), 85-102

Araştırma Makalesi

SELÇUK ÜNİVERSİTESİ

Geliş (Recieved) :29/05/2020
Kabul (Accepted) :12/09/2020

Sayma Verisi Modelleri Üzerine Bir Karşılaştırma: E- Ticarette Yaşanan Sorunlar Türkiye Örneği

Duygu KILIÇ[1*], Hülya BAYRAK[2]

*Sorumlu yazar: **duygukilic4@gmail.com**

[1]Gazi Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Bölümü, ANKARA
Orcid No: 0000-0002-3972-6648 / duygukilic4@gmail.com
[2]Gazi Üniversitesi, Fen Fakültesi, İstatistik Bölümü, ANKARA
Orcid No: 0000-0001-5666-4250 / hbayrak@gazi.edu.tr

Öz: Bu çalışmada Türkiye için e-ticarette yaşanan sorun sayısına etki eden faktörlerin belirlenmesi amaçlanmıştır. Bu amaç doğrultusunda sayma veri modellerinden yararlanılmıştır. Uygulamada 2019 TÜİK hanehalkı bilişim teknolojileri kullanım anketinde yer alan sorun sayısı verilerine Poisson (P), negatif binom (NB), sıfır yığılmalı Poisson (ZIP), sıfır yığılmalı negatif binom (ZINB), Poisson Hurdle (PH) ve negatif binom Hurdle (NBH) regresyon modelleri uygulanmıştır. Bu modellerden hangi modelin veri setini daha iyi temsil ettiği Akaike Bilgi Kriteri, log olabilirlik, Vuong, Rootogram uyum iyiliği testleri kullanılarak karar verilmiştir. Analiz sonucuna göre ZINB modelinin tercih edilmesi gerektiği görülmüştür. Ayrıca ZINB modeline ait parametreler incelenmiş ve yorumlanmıştır.

Anahtar Kelimeler: Sayma Veri Modelleri, Sıfır Yığılmalı Modeller, Hurdle Modeller, E-Ticaret

A Comparasion on Count Data Models: Example of Problems That Occured in E-Commerce Over the Turkey

Abstract: This study aimed at determining the number of problems affecting e-commerce factor for Turkey. For this purpose, count data models were used. Poisson (P), negative binomial (NB), zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), Poisson Hurdle (PH) and negative binomial Hurdle (NBH) regression models have been applied. It has been decided by using Akaike Information Criteria, log likelihood, Vuong, Rootogram goodness of fit tests which model represents the data set better. According to the results of analysis, it was seen that ZINB model should be preferred. In addition, the parameters of the ZINB model were examined and interpreted.

Keywords: Count Data, Zero Inflated Models, Hurdle Models, E-Commerce

1. Giriş

İstatistik belirli bir amaç veya olay için veri toplamada, toplanan verileri özetlemede, bu verilerle güvenilir analizler yapmada ve bu analiz sonuçlarını yorumlamada sıklıkla kullanılan bir bilim dalıdır. Önemli olan istenilen amaç için en doğru istatistiksel yöntemi kullanmak ve bu yöntemi yorumlamaktır. Bu nedenle istatistiksel analizlerde önemsenmesi gereken konulardan bir tanesi veri yapısına uygun modellerin seçilmesidir. Çeşitli veri tipleri vardır (kesikli, sürekli, nominal gibi). Bu veri tipleri arasında en çok kullanılanlardan biri kesikli verilerdir. Kesikli veriler sayılabilen verilerdir. Bu veriler sayma veri modelleri kullanarak analiz edilebilir.

Sayma veri modellerine başlamadan önce sayma verilerinin ne anlama geldiğini anlamak önemlidir. "Sayma" kelimesi genellikle birimleri, öğeleri veya olayları numaralandırmak için kullanılır (Hilbe, 2014). Bir yolda belirli bir zamanda meydana gelen kazaların sayısı düşünüldüğünde bu sayı gözlemlenebilmektedir. Sayma verileri ise, sayma yoluyla ifade edilebilen olaylar veya öğeler hakkında yapılan gözlemleri ifade etmek için kullanılır. İstatistikte sayma verileri, sıfır ve sıfırdan daha büyük değerler alabilen negatif olmayan tamsayı değerlerine sahip gözlemleri ifade etmektedir. Bu veriler kesikli bir dağılım özelliği gösterirler. Sayma verileri aktüerya, ekonometri, biyoistatistik, eğitim, sağlık gibi birçok farklı alanda kullanılır.

Sayma verilerinin modellenmesinde yanıt değişkenin geldiği dağılım önemlidir. Genellikle bu verilerde yanıt değişkeni kesikli dağılım ailesinden olan Poisson dağılımından ya da negatif binom dağılımından gelmektedir. Bu nedenle de verilerin modellenmesinde Poisson regresyon (PR) ve negatif binom regresyon (NBR) çoğunlukla tercih edilmektedir. Poisson regresyonda ortalamanın varyansa eşit olduğu bilinmektedir. Ancak günlük hayatta bu varsayıma her zaman rastlamak mümkün olmamaktadır. Varyansın ortalamaya eşit olmadığı durumlarda genellikle aşırı yayılım (overdispersion) görülmektedir. Böyle verilerin analizinde Negatif binom regresyon kullanılmaktadır. Aşırı yayılım sayma verilerinde sıklıkla rastlanan bir durumdur ve varyansın ortalamadan büyük olması şeklinde tanımlanır (Winkelman, 2000). Sayma verileri olduğu durumda dağılım genellikle sağa çarpıktır. Bu nedenle normallik varsayımı sağlanamadığından klasik regresyon yöntemleri kullanılmaz (Pittman ve ark., 2018). Eğer klasik yöntemler kullanılırsa yanlı parametre tahminleri elde edilir.

Sayma verilerinin analizinde dikkat edilmesi gereken bir başka konu ise, yanıt değişkeninin içerdiği sıfırların yoğunluğudur. Veri setinde beklenenden fazla sayıda sıfır değerin olması sıfır yığılma (zero inflation) olarak tanımlanmaktadır. Böyle veri setlerinin sıfırları göz önünde bulunduran sıfır yığılmalı modeller (zero inflated models) ile analiz edilmesi daha uygundur (Ridout ve ark., 2001). Sıfır yığılmalı modeller ekonometri, demografi, tıp, halk sağlığı, epidemiyoloji, biyoloji gibi farklı alanlarda kullanılmaktadır. Sıfır yığılmalı verilerin analizinde uygun yöntemlerin kullanılmaması, yanlı parametre tahminlerinin elde edilmesi, standart hataların küçülmesi ve tutarsız sonuçların elde edilmesine neden olabilir (Miller, 2007). Sıfır yığılmalı sayma veri modellerinin başlıcaları; sıfır yığılmalı poisson regresyon (zero inflated regression-

ZIP) ve sıfır yığılmalı negatif binom regresyon (zero negative binomial regression-ZINB) modelidir.

Sayma verilerinin analizinde Hurdle modelleri kullanmak da mümkündür. Hurdle model, veri setindeki sıfırların olasılığını modellemeyen ve pozitif sonuç üzerinde koşullanan sıralı bir dağılıma dönüşen kesilmiş (truncated) Poisson ya da negatif binom dağılımı kullanan iki bileşenli bir modeldir (Açıkyürek, 2016). Bu model, veriden sıfır olmayan değerleri tamamen ayırır ve sıfırları dikkate alır. Sıfır yığılmalı modeller Hurdle modele benzer; ancak, sıfır yığılmalı modelde sıfır olmayanlar ile sıfırların birlikte analiz edilebilmesine izin verir (Fang, 2013). Hurdle regresyon modeli iki aşamalı bir süreç içerir. Bu amaçla yanıt değişkeninin aldığı değerin sıfır veya sıfırdan farklı olma durumuna göre farklı hesaplamalar yapılır. Bu amaçla yanıt değişkenine sıfır ve pozitif değerlere karşılık olarak ikili yanıt muamelesi yapılır. Yanıt değişkeninin sıfır değerini aldığı durumlara karşılık bir olasılık tanımlar ve iki aşamalı sürecin ilk aşamasıdır. Sıfırdan farklı değerler için ise kesilmiş (truncated) dağılım kullanır ve bu da sürecinde ikinci aşamasıdır. Hurdle modelleri aşırı yayılım (over-dispersion) ve eksik yayılım (under-dispersion) durumunda da esnek ve kullanılabilirdir (Yıldırım, 2019).

Sayma verilerinin modellenmesinde bu sayılan yöntemler dışında genelleştirilmiş Poisson regresyon, kesilmiş (truncated) regresyon modelleri, sansürlü (censored) regresyon modelleri ve mixed regresyon modelleri gibi yöntemlerde kullanılmaktadır.

Lambert (1992), yaptığı çalışmada sıfır yığılmalı sayma verilerini modellemek için Sıfır yığılmalı Poisson Modellerini geliştirmiştir. Bu modeli, yanıtı bir örnek ünitedeki hatalı ürünlerin sayısı olan kalite kontrol çalışmasından toplanan verilere uygulamıştır. Daha sonra Greene (1994), tarafından aşırı yayılım gösteren veriler için negatif binom modelinin genişletilmiş bir formu olan sıfır yığılmalı negatif binom model tanımlanmıştır. Carrivick ve ark. (2003), işçilerin yaralanma sayısı üzerindeki etkisinin incelenmesinde sıfır yığılmalı Poisson modeli kullanmışlardır. Martin ve ark. (2005), sıfır yığılmalı modeller ve Hurdle modeller kullanılarak, belirli bir kuş türünün sayısının modellenmesi üzerine çalışma yapmışlardır. Yang ve ark. (2009), sıfır yığılmalı Poisson model için aşırı yayılımları test etmede kullanılan modeller üzerine çalışma sunmuşlardır. Kaya ve Yeşilova (2012), çalışmalarında Yüzüncü Yıl Üniversitesi e-posta trafiğini sıfır yığılmalı regresyon modelleri kullanarak incelemişlerdir. Asrul ve Naingb (2012), çalışmalarında AIDS hastalarının ölüm oranlarını yaş üzerinden modellemede cinsiyet, milliyet, ırk, medeni durum, meslek ve taşıyıcılık şekli değişkenlerinin etkisini sıfır değer ağırlıklı regresyon modelleri

kullanarak ortaya koymuşlardır. Sharma ve Landge (2013), çalışmalarında Hindistan'da yer alan bir karayolu üzerindeki kaza sayılarını sıfır yığılmalı modeller kullanarak incelemişlerdir. Kong ve ark. (2014), diş çürüğü ile florür uygulaması arasındaki ilişkiyi sıfır yığılmalı modellerden ZINB model ile modellenmiştir. Wang ve ark. (2015), Almanya'da sağlık hizmetleri çalışmalarını sıfır yığılmalı negatif binom model kullanarak incelemişlerdir. Beaujean ve Morgan (2016), eğitim alanından kullandıkları veri ile sıfır değer ağırlıklı veri seti için uygun model seçimi hakkında çalışma sunmuşlardır. Tüzen ve Erbas (2017), kamu sağlığı ve madde kullanımı konularında sıfır yığılmalı modeller ve Hurdle modellerin kullanılması ve karşılaştırılmasına konusunda çalışma yapmışlardır. Altun (2018), yapmış olduğu çalışmada sıfır yığılmalı modellere yeni bir yaklaşım olarak sıfır yığılmalı Poisson-Lindley modelini önermiştir. Karaca ve Olmuş (2018), sıfır yığılmalı verilerin analizinde sıfır yığılmalı regresyon modellerin incelenmesi üzerine çalışmışlardır. Kim ve ark. (2019), yapmış oldukları çalışmalarında Güney Kore'de iklim değişikliği sonucu aşırı sıcağa maruz kalan kişilerde yaşanan ölüm sayısını tahmin etmek için sıfır yığılmalı regresyon modelleri kullanmışlardır. Özen (2020), yapmış olduğu çalışmasında Mersin ili için seçili kavşaklarda gerçekleşen trafik kazalarının sayısına etki eden faktörleri sayma veri modelleri kullanarak incelemiştir.

2. Materyal ve Metot

Hanehalkı Bilişim Teknolojileri araştırması, hanelerde ve bireylerde sahip olunan bilgi ve iletişim teknolojileri ile bunların kullanımları hakkında bilgi derlemek amacıyla uygulanmakta olup, söz konusu teknolojilerin kullanımı hakkında bilgi veren temel veri kaynağıdır. Bu araştırma ile hanelerde bulunan bilgi ve iletişim teknolojileri, hanehalklarının veya bireylerin bilgi ve iletişim teknolojilerine erişimi ve kullanımı, e-ticaret, e-devlet uygulamaları, bilişim güvenliği gibi alanlarda veri derlenmektedir. Bu başlıklar ihtiyaca göre değişebilmektedir. Araştırma her yıl Nisan ayında gerçekleştirilmektedir (TÜİK). Bu çalışmada sayma veri modellerinin performanslarına ilişkin farklı durumlar gerçek bir veri seti ile ele alınmaya çalışılmıştır. Türkiye' de e- ticarette yaşanan sorun sayısı ve bu sorun sayısına etki eden faktörleri incelemek için 2019 yılı TÜİK Hanehalkı Bilişim Araştırması mikro veri seti kullanılmıştır. Sayma veri modellerinden poisson regresyon, negatif binom regresyon, sıfır yığılmalı poisson regresyon, sıfır yığılmalı negatif binom regresyon, poisson Hurdle ve negatif binom Hurdle modelleri dikkate alınmıştır. Model karşılaştırılmasında akaike bilgi kriteri

(AIC), log olabilirlik (LL), Vuong istatistiği ve rootogram kullanılmıştır.

2.1. Sayma Veri Modelleri

2.1.1. Poisson regresyon modeli (PR)

Poisson regresyon (PR), sayma verilerinin modellenmesinde en sık kullanılan, en basit ve en temel yöntemlerden biridir. Bu modelde yanıt değişkeni Poisson dağılımından gelmektedir. Poisson dağılımı tek parametreli bir dağılımdır (Winkelman, 2000). Poisson dağılımında log bağlantı fonksiyonu kullanıldığından Loglineer model olarakta adlandırılmaktadır. (Agresti, 2002). Poisson regresyon tıp, biyoloji, demografi ve ekonometri gibi birçok alanda kullanılmaktadır. Poisson regresyon belirli bir zaman aralığında meydana gelmiş olan olaylar ile bu olaylar için belirlenen bağımsız değişkenler arasında bir bağlantı kurulmak istendiği durumlarda kullanılmaktadır (Yıldırım, 2019).

Ortalama ile varyansın birbirine eşit olması bu modelin en belirgin özelliğidir. Model Eş. 1'de verilmiştir.

$$P(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!},\ y= 0, 1, 2 \ldots. \qquad (1)$$

Burada μ ortalama, y ise yanıt değişkenini ifade etmektedir. Ortalama ve varyans, E(Y)= Var(Y)= μ dür.

2.1.2. Negatif binom regresyon modeli (NBR)

Poisson regresyonun özel bir halidir. Poisson regresyonda ortalama ve varyans birbirine eşit iken bu durum uygulamada her zaman mümkün olmamaktadır. Veri kümesinde aşırı yayılım olması durumunda, yayılımı dikkate alan negatif binom regresyonun kullanılması daha uygun olmaktadır (Sileshi, 2008). Modele aşırı yayılımdan kaynaklanan etki için yeni bir parametre ekleyerek analiz yapılır. Eğer aşırı yayılım durumu gözardı edilirse yanlı parametre tahminleri, tutarsız sonuçlar elde edilir. Model Eş. 2'de verilmiştir.

$$P(Y = y) = \frac{\Gamma\left(Y+\frac{1}{k}\right)}{\Gamma(Y+1)\Gamma\left(\frac{1}{k}\right)}\left(\frac{1}{1+k\mu}\right)^{1/k}\left(\frac{k\mu}{1+k\mu}\right)^{Y} \qquad (2)$$

Burada k aşırı yayılım parametresidir. Ortalama ve varyans ise E(Y)= μ ve Var(Y)= μ(1+kμ) şeklinde ifade edilir.

2.1.3. Sıfır yığılmalı poisson regresyon modeli (ZIP)

ZIP, veri setindeki yanıt değişkeninin Poisson dağılımdan geldiği ve bu değişkenin beklenenden daha fazla sıfır içerdiği durumlarda kullanılır. Sıfır yığılma durumunda ortaya çıkan sıfır değerler elde ediliş durumlarına göre yapısal sıfır ve örnekleme sıfırı olarak iki şekilde adlandırılmaktadır. Değerlendirilen veri setinde elde edilme ihtimali olmadığından yani gözlem yapmanın imkân dâhilinde olmadığı durumlarda hücrenin sıfır değerini

almasına yapısal sıfır; elde edilmesi mümkün olduğu halde veri setinde gözlenmeyen, bu nedenle sıfır değer alan değerlere de örnekleme sıfırı denir (Karaca, 2018).

Bağımlı değişkeninin iki farklı tip veriden oluştuğunu varsayılmaktadır. Bunlardan birincisi, sıfır değerlerini de içerebilecek Poisson dağılımlı veri grubu olurken, ikinci grup ise daima sıfır değerlerini içeren gruptur (Cameron ve Trivedi, 2013). Bu nedenle model iki parçalı şekilde ifade edilir. Modelin log dönüşüm fonksiyonu ile modellenen kısmı poisson dağılımından gelen pozitif sayıları modellemede kullanılırken, logit kısmı veri setindeki sıfırları modellemede kullanılır (Peng, 2013). Model Eş. 3'de verilmiştir.

$$P(Y=y)=f(x)=\begin{cases} w+(1-w)e^{-\mu}, & y=0 \\ (1-w)\frac{e^{-\mu}\mu^{y}}{y!}, & y\geq 1, \quad 0\leq w\leq 1 \end{cases} \tag{3}$$

Burada *w*, sıfır yığılma olasılığını göstermektedir.

2.1.4. Sıfır yığılmalı negatif binom regresyon (ZINB)

ZINB, veri setinde çok sayıda sıfır olduğunda ve veride aşırı yayılım durumu olduğunda kullanılmaktadır. Bu regresyon modeli, Poisson regresyonun açıklayamadığı aşırı yayılımı ve sıfır değer ağırlığını modellemek için geliştirilmiş ve bağımlı değişkendeki heterojenliği açıklamak için uygun bir regresyon yöntemidir (Zhuo ve ark., 2008). ZIP modelde olduğu gibi sıfır değeri alan gözlemler ile sıfır olmayan gözlemler ayrı olarak modellenir. Ancak ZIP dan farklı olarak bu regresyon modelinde sıfır olmayan gözlemler negatif binom regresyonu ile modellenmektedir. Model Eş.4'de verilmiştir.

$$P(Y=y)=\begin{cases} w+(1-w)\left(\frac{1}{1+k\mu}\right)^{\frac{1}{k}}, & y=0 \\ (1-w)\frac{\Gamma\left(y+\frac{1}{k}\right)}{\Gamma(y+1)\Gamma\left(\frac{1}{k}\right)}\left(\frac{1}{1+k\mu}\right)^{\frac{1}{k}}\left(\frac{k\mu}{1+k\mu}\right)^{y}, & y\geq 1 \end{cases} \tag{4}$$

Burada *k* sıfıra yaklaştığında model ZIP modeline dönüşmektedir. Bu da bu iki modelin iç içe (nested) modeller olduğunu gösterir.

2.1.5. Poisson hurdle regresyon modeli (PH)

Sıfır yığılmalı sayma modellerine alternatif bir modeldir. Pozitif değerler alan bölüm Poisson dağılımından gelmekte ise bu

yöntem kullanılmaktadır. Model Eş. 5'de verilmiştir.

$$P(Y=y)=\begin{cases} w, & y=0 \\ (1-w)\frac{e^{-\mu}}{(1-e^{-\mu})\Gamma(Y+1)}, & y\geq 1 \end{cases} \quad (5)$$

Burada *w*, sıfır yığılma olasılığını ifade eder.

2.1.6. Negatif binom hurdle regresyon modeli (NBH)

Diğer sayma veri modellerinde olduğu gibi bu modelde aşırı yayılım olması durumunda tercih edilir. Pozitif değer alan bölüm negatif binom dağılımından gelmekte ise bu yöntem kullanılmaktadır. Model Eş. 6'da verilmiştir.

$$P(Y=y)=\begin{cases} w, & y=0 \\ (1-w)\frac{\Gamma\left(y+\frac{1}{k}\right)^{w}}{[1-(\frac{1}{1+k\mu})^{\frac{1}{k}}]\Gamma(y+1)\Gamma(\frac{1}{k})}(\frac{1}{1+k\mu})^{\frac{1}{k}}(\frac{k\mu}{1+k\mu})^{y}, & y\geq 1 \end{cases} \quad (6)$$

.
.

Burada *k* yayılım parametresidir.

3. Model Seçimi

Sayma veri modellerinde hangi modelin daha uygun olduğunu belirlemek amacıyla çeşitli testlerden yararlanılabilmektedir. Bu çalışmada Akaike Bilgi Kriteri (AIC), log olabilirlik (LL) değeri ve Vuong istatistiği kullanılmıştır. AIC model denklemi;

AIC= -2lnL+ 2p şeklindedir. Burada lnL logaritmik olabilirlik değeri, p parametre sayısıdır.

Bir modelin iyi olduğu yorumu AIC değeri en küçük olduğunda yada LLdeğeri en büyük olduğunda yapılabilir.

Vuong testi, iç içe geçmeyen (non-nested) modellerin karşılaştırılmasında kullanılan hipotez testlerinden biridir (Tüzel, 2011). İç içe model karşılaştırmalarının dışında olası ikili modeller de Vuong testiyle karşılaştırılabilmektedir. Bu sayede sıfır yığılmalı modellerde hangi modellerin uygun olabileceği belirlenebilmektedir.

$p_1(.)$ ve $p_2(.)$ Olasılık yoğunluk fonksiyonları olmak üzere Vuong testi için denklemler Eş.7 ve Eş.8'de şekilde ifade edilmiştir.

$$V=\frac{\bar{m}\sqrt{n}}{sd(m)} \quad (7)$$

Burada $\bar{m}$, m_i nin ortalamasını, *sd(m);* standart sapmayı ve *n* ise örnek çapını temsil etmektedir. m_i ise şu şekilde ifade edilir:

$$m_i = ln\left(\frac{p_{1i}(y_i)}{p_{2i}(y_i)}\right) \quad (8)$$

Vuong test istatistiği standart normal dağılımlıdır. Vuong test değerinin yorumlanmasına örnek olarak, 0.05 anlamlılık düzeyi için, *V* değeri 1.96'dan büyükse, ilk model gerçek modele "daha yakındır" şeklinde bir yorum yapılabilirken eğer *V* değeri -1.96'dan küçükse ikinci model gerçek modele "daha yakındır" yorumu yapılabilir. Hesaplanan değer (-1.96; 1.96) arasında değil ise hiçbir model gerçek modele "daha yakın" değildir yada birinci veya ikinci modeli kullanma arasında fark yoktur yorumu yapılır (İsmail ve Zamani, 2013).

Modelleri karşılaştırmada bir diğer yöntem olarak Rootogram kullanılabilir. Tukey (1977) çalışmasında ilişkili bir grafik aracı olan rootogram'ı kullanmış ve başlangıçta tek değişkenli dağılımların uyumluluğunu değerlendirmiştir. Kleiber ve Zeile (2016), Rootogram'ı regresyon modelleri için genişletmiştir ve bu grafiğin özellikle sayma veri modellerinde aşırı dağılım ve/veya sıfır yığılma gibi sorunların anlaşılmasında yararlı olduğunu öne sürmüşlerdir. Rootogram, gözlenen frekanslar için histogram benzeri dikdörtgenler veya çubuklar ve frekanslar için bir kare kök ölçeğinde bir eğri çizerek gözlenen ve beklenen değerleri grafiksel olarak karşılaştırır. Üç tür Rootogram vardır: sarkık (suspended), asılı (hanging) ve askıya alınmış (standing). Bu çalışmada Rootogramın asılı versiyonu kullanılmıştır.

4. Sonuçlar ve Tartışma

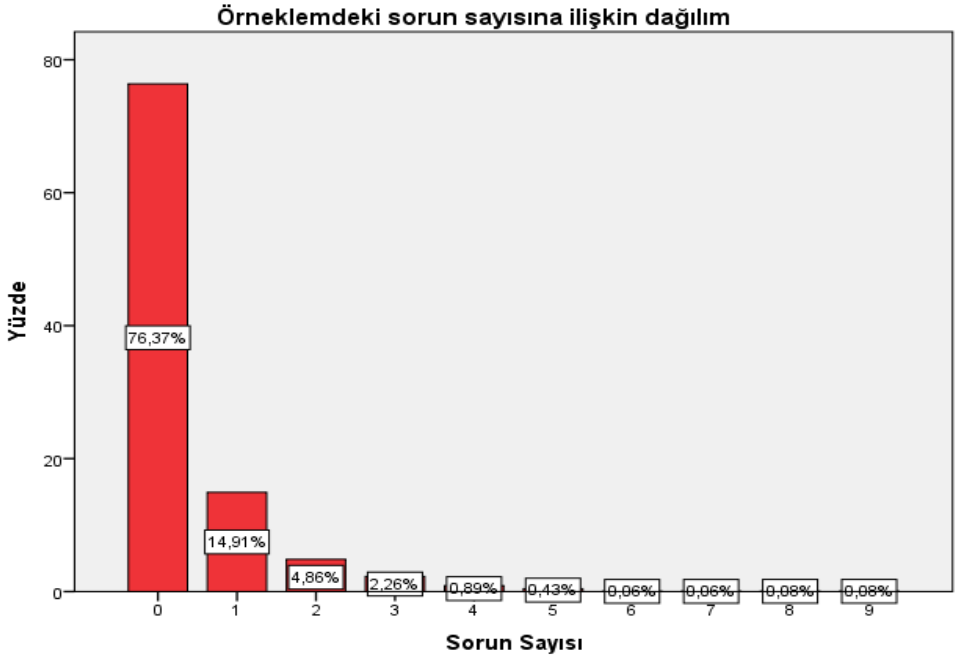
Bu çalışmada TÜİK (Türkiye İstatistik Kurumu) üzerinden alınan Hanehalkı Bilişim Teknolojileri Kullanımı Araştırması Mikro Veri Seti (2019) kullanılmıştır. TÜİK, 2004 yılından itibaren yıllık periyotlarla bu araştırmayı yapmaktadır. Kurumun bu araştırmayı yapma amacı; bilgi toplumu ölçütlerinin belirlenmesi ve ilgili istatistiklerin üretilmesidir (TÜİK). Anket toplamda 16010 haneye yapılmıştır. Araştırma metodolojisi gereği 16-74 yaş arasındaki bireyleri kapsamaktadır. Bu anket içinde e-devlet kullanımı, e-ticaret kullanımı ve internet kullanımı gibi başlıklar mevcuttur. Bu çalışmada e-ticaret bölümü kullanılmıştır. Veriler, belirlenen örnekleme yöntemine göre seçilen hanehalklarından derlenmektedir. Bağımlı değişken olarak son 12 ay içinde internet üzerinden mal ve hizmet siparişi verirken ya da satın alırken karşılaşılan sorun sayısı alınmıştır. Bu değişken ile ilişkisi olduğu düşünülen son 12 ay içinde internet üzerinden alınan mal veya hizmet sayısı, son 12 ay içerisinde yurtiçindeki satıcılardan, Avrupa Birliği ülkelerindeki satıcılardan ve satıcının ülkesini bilmediği durumlarda ürün

veya hizmet satın alma durumu ve internette yapılan işlem sayısı değişkenleri bağımsız değişken olarak seçilmiştir. Değişken seçiminde literatürde yapılan çalışmalar göz önüne alınmıştır. Analiz R Studio programı kullanılarak yapılmıştır. Tablo 1 de analizde kullanılan bağımsız değişkenler verilmiştir.

Tablo 1. Veri setinde yer alan bağımsız değişkenler

Değişken Adı	Açıklaması	Değer Aralığı
İşlem sayısı	İnternet üzerinden kaç çeşit işlem yaptınız?	Min:0 Maks: 14
Çeşit sayısı	İnternet üzerinden kaç çeşit mal ve hizmet aldınız?	Min:1 Maks:15
Yurtiçi	Yurtiçindeki satıcılardan mal veya hizmet satın aldınız mı?	Evet:1 Hayır:2
AB ülkeleri	AB ülkelerindeki satıcılardan mal veya hizmet satın aldınız mı?	Evet:1 Hayır:2
Bilinmiyor	Ülkesini bilmediğiniz bir satıcıdan mal veya hizmet aldınız mı?	Evet:1 Hayır:2

İlgili veri setinde kullanılan bağımlı değişkenin dağılımı Şekil 1'de verilmiştir.



Şekil 1. E-ticarette yaşanan sorun sayısının dağılımı

Şekil 1'e göre hiç sorun yaşamayan grup örneklem içinde 0.76 lık bir paya sahiptir. Buna karşın 1 sorunla karşılaşan birey oranı %15, 2 sorunla karşılaşan birey oranı %5 şeklindedir. Sorun sayısı artıkça yüzdesel bir düşüş görülmektedir.

Sorun sayısının dağılımı			
		Sayısal Değeri	**Yüzdesi**
Aldığı Değerler	0	6038	,6
	1	1179	,1
	2	384	,0
	3	179	,0
	4	70	,0
	5	34	,0
	6	5	,0
	7	5	,0
	8	6	,0
	9	6	,0

Tablo 2'ye bakıldığında verilerin %76'sının (6038 gözlemin) 0 değerini aldığı görülmüştür. Bu durum veri setinin sıfır yığılmalı modeller için uygun olduğunu göstermektedir. Modeldeki sıfır sayısına bakılarak aslında e-ticaret üzerinden mal ve hizmet satın alan kişilerin yarısından fazlasının herhangi bir sorunla karşılaşmadığı görülmektedir.

Yukarıda bahsedilen 6 model veri setine uygulanmıştır. Sonuçlar için ilk olarak model uyum iyiliğine bakılmıştır. Bu bağlamda AIC, LL, Vuong istatistiği ve Rootogram kullanılmıştır. Daha sonra bu kriterlere göre seçilen en iyi model belirlenmiş ve parametre tahminleri verilmiştir.

4.1. Model Seçimi

Altı modelin uygunluğunu karşılaştırmak için ilk olarak AIC ve LL değerine bakılmıştır. AIC değeri en küçük olan, LL değeri ise en büyük olan model tercih edilir. Tablo 3'e bakıldığında karşılaştırma sonucunda en düşük AIC ve en büyük LL değerine göre, ZINB modelinin verilere diğer modellerden daha iyi uyduğu görülmüştür. En büyük AIC ve en küçük LL değerinin gözlendiği Poisson regresyon (PR) modeli ise diğer modeller arasında en az uyum gösteren model olmuştur.

Tablo 4. Modellere ilişkin AIC ve LL değerleri

	AIC	**LL**
PR	13394.1	-6691.052
NB	12381.83	-6183.917
ZIP	12432.39	-6204.196
ZINB	12318.51	-6146.254
PH	12433.62	-6204.808
NBH	12327.19	-6150.593

Vuong istatistiği ile karşılaştırma yapılmak istendiğinde ise her model karşılaştırması ilgili model çiftinde model 1'e karşı model 2 şeklinde yapılır. Pozitif test istatistiği model 1'in model 2'ye tercih edildiğini, negatif test istatistiği model 2'nin model 1'e tercih edildiğini göstermektedir. Ayrıca Vuong testi için seçilen çiftlerin iç içe olmayan (nonnested) çiftler olması

gerekmektedir. Karşılaştırılan model çiftleri Tablo 4'de verildiği gibidir.

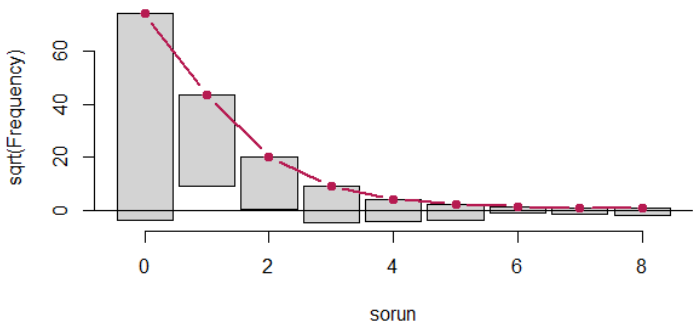
Tablo 4. Vuong istatistikleri

(1. model-2. model)	*V* istatistiği	P değeri
P- ZIP	-11.87899	2.22e-16
NB- ZINB	-4.723123	1.1612e-06
ZIP-PH	1.111958	0.13308
NB-NBH	-4.4347362	4.6093e-06
P-PH	-11.87517	2.22e-16
ZINB-NBH	1.265171	0.10291

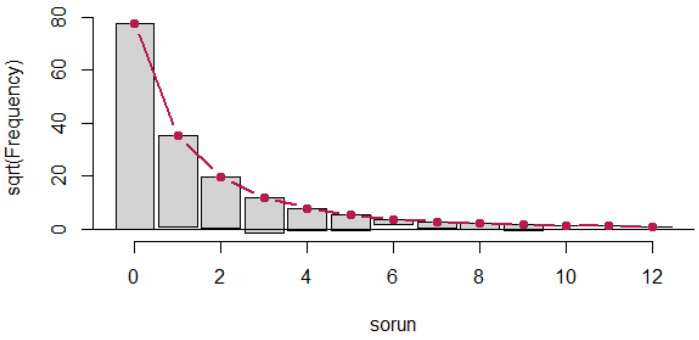
Vuong testi sonuçlarına göre P – ZIP model karşılaştırması değerlerine göre *V* istatistik değerinin -1.96 değerinden küçük olduğu belirlenmiştir. Model 2 olarak tanımlanan ZIP modelinin PR modeline göre veri setini açıklamada daha uygun model olduğu sonucuna varılmıştır. Aynı durum NB – ZINB model karşılaştırmasında da görülmektedir. *V* istatistiği -1.96 değerinden küçük olarak hesaplanmıştır. Böylece model 2'nin yani ZINB modelin NB modelinden daha üstün olduğu yorumu yapılabilir.

Bu altı modeli görsel olarak karşılaştırılmak için Rootogram grafiğinden yararlanılmıştır. Bu modeller için elde edilen grafikler Şekil 2'de sıralanmıştır.

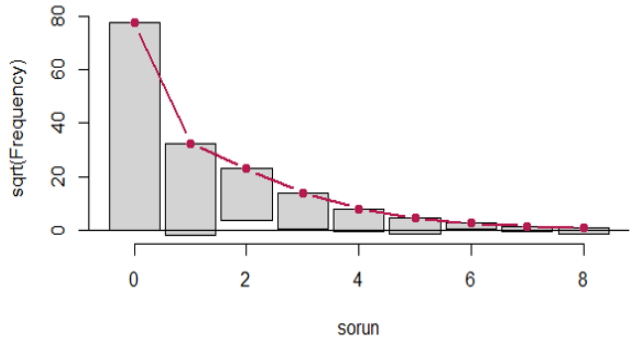
P regresyon modeli için:



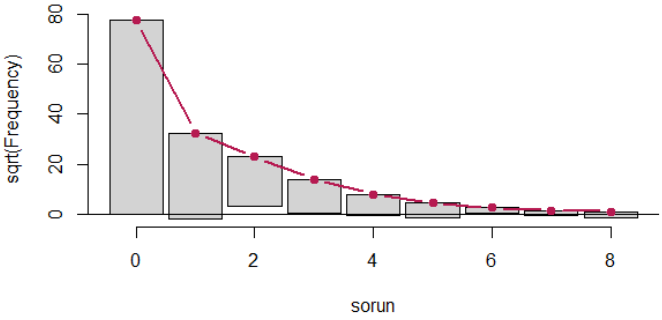
NB regresyon modeli:



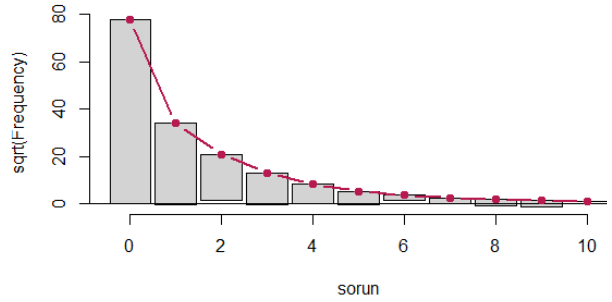
ZIP regresyon modeli:



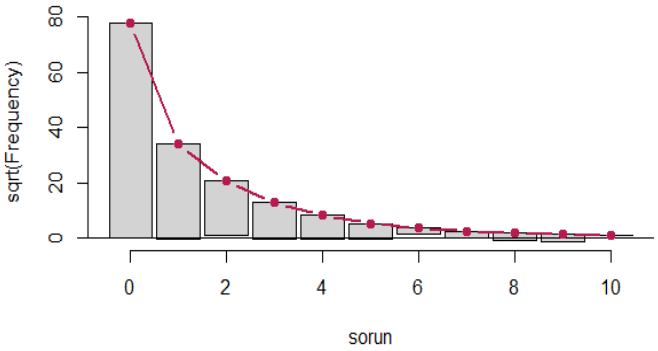
PH regresyon modeli:



ZINB regresyon modeli:



NBH regresyon modeli



.

Şekil 2. Rootogram grafikleri

Şekil 2'ye göre barlar (ya da çubuklar) tahmin edilen ve gözlenen sorun sayılarının kareköklerinin farkını göstermektedir. Çizgiler ise tahmin edilen sorun sayısının karekökünü ifade etmektedir. Barlar x eksenine ne kadar yakınsa model o kadar iyidir şeklinde yorum yapmak mümkündür. Buna göre Poisson modeline ait grafikte barların sıfıra diğer modellere göre daha uzakta olduğunu görülmektedir. Poisson modeline ilişkin AIC ve LL değerlerine de bakıldığında modelin veri setine uyumunda en kötü model olduğu görülmektedir. ZINB ve NBH grafiklerinin oldukça benzer oldukları görülmektedir. Yine AIC ve LL değerlerine bakıldığında bu iki modelin veri setine uygunluğunun yakın olduğu söylenebilir.

4.2. Parametre Yorumu

Yapılan uyum iyiliği testlerine göre ZINBmodelinin veri setini temsil etmekte en iyi model olduğu görülmüştür. Seçilen ZINB modeline ilişkin parametre tahmin değerleri aşağıdaki tabloda verilmiştir.

Tablo 5. ZINB modeline ait parametre tahminleri

Log Kısım	Tahmin Değeri	Standart Hata	Z değeri	P değeri	e^{β}
Sabit Terim	-0.49212	0.25008	-1.968	0.049084 *	0,611329
İşlem Sayısı	0.03743	0.01771	2.114	0.034513 *	1,038139
Çeşit Sayısı	0.08182	0.01346	6.081	0.00001209*	1,08526
Yurtiçi	0.12983	0.18953	0.685	0.493326	1,138635
AB Ülkeleri	-0.16151	0.11596	-1.393	0.163697	0,850858
Ülke bilinmiyor	-0.50993	0.1346	-3.882	0.000104*	0,600538

Logit Kısım	Tahmin Değeri	Standart Hata	Z değeri	P değeri	e^{β}
Sabit Terim	-0.10390	0.63701	-0.163	0.8704	0,901315
İşlem Sayısı	0.07647	0.04128	1.853	0.0639	1,07947
Çeşit Sayısı	-0.38138	0.06829	-5.584	2.35e-08 *	0,682918
Yurtiçi	-0.39119	0.43828	-0.893	0.3721	0,676252
AB Ülkeleri	0.41017	0.43915	0.934	0.3503	1,507074
Ülke bilinmiyor	0.07647	0.04128	1.853	0.0639	1,07947

Log kısım için sabit terim, işlem sayısı, çeşit sayısı ve satıcının ülkesinin bilinmemesi değişkenleri anlamlı çıkmıştır ($p< 0.05$). Log kısım için elde edilen regresyon denklemi μ= exp(-0.49+ 0.038 işlem sayısı+ 0.082 Çeşit sayısı+ 0.13 Yurtiçi- 0.16 ABülkeleri- 0.51Ülke bilinmiyor) olarak elde edilmiştir. İşlem sayısındaki değişiklik sorun sayısını ($e^{\wedge 0.04}$= 1.038~%4) %4 değiştirmektedir. Yani kişilerin internet üzerinden yaptıkları işlem sayısı artıkça yaşadıkları sorun sayısı da artmaktadır. Çeşit sayısındaki değişiklik sorun sayısını %8 değiştirmektedir. Kişiler e- ticaret üzerinden daha fazla ürün aldıkça yaşadıkları sorun sayısı da artmaktadır. Sorun sayısında kişinin ürünü aldığı satıcının ülkesini bilmesi durumu bilmemesi durumuna göre ($e^{\wedge -0.51}$=0.6) %40 azaltmaktadır.

Logit kısım için ise çeşit sayısı istatistiksel olarak anlamlı çıkan tek değişkendir. Logit kısım için elde edilen regresyon denklemi $\left(\frac{\pi}{1-\pi}\right)$= exp (-0.104+ 0.077 İşlem sayısı- 0.381 Çeşit sayısı -0.391

yurtiçi+ 0.41 AB ülkeleri+ 0.076 Ülke bilinmiyor) şeklindedir. Çeşit sayısındaki değişiklik sorun sayısını %32 değiştirmektedir. Buna göre alınan ürünlerinin çeşitliliğinin sayısı sorun yaşanmasında etkili olmaktadır. Bunun arkasında birçok sebep yatıyor olabilir. Yaşanan sorunların başlıcaları, teslimatın yavaşlığı, teknik arıza, hatalı ürün gönderimi ve dolandırıcılıktır.

Literatürde e-ticaret üzerine yapılan çalışmalar yirminci yüzyılın son çeyreğinde başlamış ve günümüzde de hız kesmeden devam etmektedir.

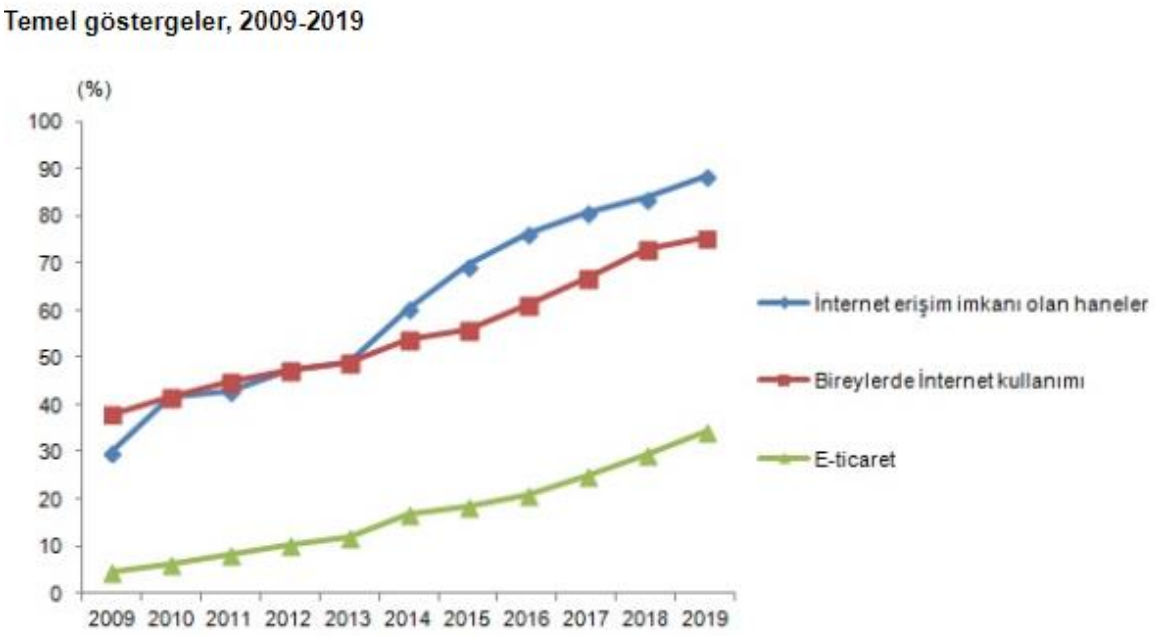
Dünyanın gelişmiş ekonomilerine bakıldığında son yıllarda beklentilerin ötesinde bir büyüme gerçekleştiği görülmektedir. Bunun arkasında yatan temel etkenlerden biri şüphesiz bilgi ve iletişim teknolojilerinde geldikleri üst seviye, bilgisayarla çalışma eğiliminde yaşanan gelişme ve internetin her alanda yaygın kullanılmasıdır (Canpolat, 2001). İşletmeler e-ticarette minimum sermaye ile kolay ve hızlı bir şekilde dünya genelinde daha fazla sayıda müşteriye, tedarikçiye ve uygun iş ortaklarına ulaşabilmektedir.

E-ticaretin yaygınlaşması uluslararası ticaret, eğitim, kültür ve sosyal yaşam gibi birçok alanı da etkilemektedir. E- ticaret, internet kullanımının ucuzlaması ve yaygınlaşması ile ayrıca kredi kartlarının kullanımının artması, bankacılık alanındaki yenilikler ve gelişmeler sonucunda artışını sürdürmektedir.

E- ticaret, bilgi ve iletişim teknolojilerindeki gelişmelere paralel olarak ülkelerin ekonomik ve sosyal yapılarını etkilemektedir. Ulusal pazarların sınırları, özellikle elektronik ticarete elverişli sektörlerde küreselleşmektedir. Dinamik ve sürekli büyüyen bir yapıya sahip bu pazarlarda, bilgisayar ve internet kullanım oranlarının yükselmesiyle birlikte elektronik ortamdaki tüketici sayısının da artması işletmeleri, elektronik ticaretten pay alma konusunda yeni yaklaşımlara zorlamaktadır. İnternetin ulaştığı tüm ülkelerdeki birey ve firmalar müşteri veya satıcı konumuna gelmekte, ticari işlemler fiziki çevreden soyutlanarak sanal ortama taşınmaktadır. Piyasaların, müşterileri ve satıcıların elektronik ortamlarda fiziki sınırları aşarak bir araya gelmesi, pazarın büyüklüğünü artırmakta ve elektronik ticaret için uygun ortamlar yaratmaktadır.

21.yy dan itibaren gelişen teknoloji ile tüketicilerin tüketim alışkanlıkları değişmeye başlamıştır. Gelişen teknoloji sayesinde tüketiciler istedikleri zaman istedikleri ürün ve hizmetlere kolayca hızlı bir şekilde ulaşabilmektedir. İşletmelerde gelişen bu teknoloji karşısında çağa ayak uydurmuş fiziksel mağazalara ek olarak ürün ve hizmetlerini elektronik ticaret platformları üzerinden de sunmaya başlamıştır. Günümüzde e-ticarete olan ilgi

her geçen yıl biraz daha artmaktadır. Bunun bir göstergesi olarak internet üzerinden kişisel kullanım amacıyla mal veya hizmet siparişi veren ya da satın alan 16-74 yaş grubundaki bireylerin oranının 2018 yılına göre 2019 yılında %5 arttığı TÜİK tarafından açıklanmıştır. 2018 yılı Nisan ayı ile 2019 yılı Mart aylarını kapsayan on iki aylık dönemde bu oran %34.1 iken önceki yılın aynı döneminde ise bu oran %29.3 olarak gözlenmiştir (TÜİK, 2019).



Şekil 3. TÜİK verilerine göre e-ticaret kullanımı artış grafiği (TÜİK)

Tüketiciler, elektronik ticareti güvenli internet bağlantıları ve güvenli ödeme sistemleri kullanarak mal ve hizmet satın alma süreci olarak tanımlanmaktadır.

E-ticaret sayesinde üreticilerin aracıları ortadan kaldırıp ürünlerini internet üzerinden aracısız olarak müşterilere satacağı ve böylece işlem maliyetlerini azaltabileceği düşünülmektedir. Bu durum daha düşük üretim maliyetlerine yol açacağından, piyasalara yeni girişleri teşvik edebileceği düşünülmektedir (Çilan ve Sultan, 2013).

E-ticaretin bazı avantajları ve dezavantajları vardır. Misevičiūtė (2001) yaptığı çalışmasında, avantajlarını şu şekilde sıralamıştır: kolaylık sağlama, detaylı bilgi ve öneri verme, daha düşük fiyat ve daha fazla seçenek imkanı sağlama, daha geniş bir tüketici ağına sahip olma ve çok çeşitli ürün imkanıdır. Yine Misevičiūtė (2001) dezavantajlarını ise şu şekilde sıralamıştır: dolandırıcılık, hatalı fiyatlandırma, nakliye süreçleri ve iade süreçleridir.

Bu çalışmada e-Türkiye için e-ticarette yaşanan sorun sayısına etki eden

faktörlerin belirlenmesi amaçlanmıştır. Bu amaç doğrultusunda sayma veri modellerinden bazıları kullanılmıştır. Bağımlı değişken olan e-ticarette yaşanılan sorun sayısı değişkenin ortalaması 0.39 ve varyansı 0.81 olarak bulunmuştur. Varyans ortalamadan büyük olduğundan veride aşırı yayılım olduğu görülmektedir. Ayrıca verilerin %76'sı sıfır değerini aldığından veride sıfır yığılma durumu söz konusudur. Bu bilgiler ve model seçimi kriterlerine göre veri seti için en uygun modelin ZINB modeli olduğu görülmüştür. Daha sonra ZINB modeline ilişkin parametreler yorumlanmıştır.

Sonuç olarak bu çalışmada Türkiye'de e-ticarette yaşanılan sorun sayısı incelenmiş, ankete katılan bireylerden çoğunun bir sorun yaşamadığı görülmüştür. Klasik sayma regresyon modelleri (Poisson ve Negatif binom) ve Hurdle regresyon modellerinin de parametre tahminlerine bakılmış sorun sayısına etki eden ana faktörün alınan ürün çeşitliliği olduğu görülmüştür.

Kaynaklar

Açıkyürek G (2016). Poisson regresyon ve bir uygulama (Yüksek Lisans Tezi). *Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü*, Ankara.

Agresti A (2002). Categorical data analysis (Second Edition), *New Jersey: Wiley & Sons Incorporation*.

Altun E (2018). A new zero-inflated regression model with application. *İstatistikçiler Dergisi: İstatistik ve Aktüerya* 11(2): 73-80.

Asrul AAM, Naingb NN (2012). Analysis death rate of age model with excess zeros using zero inflated negative binomial and negative binomial death rate: mortality AIDS co-infection patients, Kelantan Malaysia. *Procedia Economics and Finance* 2: 275-283.

Beaujean AA, Morgan GB (2016). Tutorial on using regression models with count outcomes using R. *Practical Assessment Research & Evaluation* 21(2): 1531-7714.

Cameron AC, Trivedi PK (2013). Regression analysis of count data (Second Edition). *New York: Cambridge University Press.*

Canpolat Ö (2001). E-ticaret ve Türkiye'deki gelişmeler. *Sanayi ve Ticaret Bakanlığı.*

Carrivick PJW, Lee AH, Yau KKW (2003). Zero-inflated poisson modeling to evaluate occupational safety interventions. *Safety Science* 41(1): 53-63.

Çilan ÇA, Sultan KUZU (2013). Kişisel e-ticaret uygulamalarının kategorik veri analizi yöntemleri ile değerlendirilmesi. *Alphanumeric Journal* 1(1): 27-32.

Fang R (2013). Zero-ınflated negative binomial (ZINB) regressıon model for over dispersed count data with excess zeros and repeated measures an application to human microbiota sequence data (Yüksek Lisans Tezi).

Greene WH (1994). Accounting for excess zeros and sample selection in poisson and negative binomial regression models. *New York University Department of Economics Working Paper* 94-10.

Hilbe JM (2014). Modelling count data (First Edition). *New York: Cambridge University Press.*

Ismail N, Zamani H (2013). Estimation of claimcount data using negative binomial, generalized poisson, zero-inflated negative binomial and zero-inflated generalized poisson regression models. *Casualty Actuarial Society E-Forum* 41(20): 1–28.

İnternet:http://www.tuik.gov.tr/PreHaberBultenleri.do?id=30574

Karaca AG (2018). Sayma verileri için regresyon modellerinin karşılaştırılması üzerine bir uygulama (Yüksek Lisans Tezi). *Gazi Üniversitesi Fen Bilimleri Enstitüsü,* Ankara.

Karaca AG, Olmuş H (2018). Sıfır değer ağırlıklı verilerin analizinde sıfır değer ağırlıklı regresyon modellerin incelenmesi. *Trakya Üniversitesi Sosyal Bilimler Dergisi* 20(2): 105–118.

Kaya Y, Yeşilova A (2012). E-posta trafiğinin sıfır değer ağırlıklı regresyon yöntemleri kullanılarak incelenmesi. *Anadolu University of Sciences & Technology-A: Applied Sciences & Engineering* 13(1): 51–63.

Kim DW, Deo RC, Park SJ, Lee JS, Lee WS (2019). Weekly heat wave death prediction model using zero-inflated regression approach. *Theoretical and Applied Climatology* 137(1-2): 823–838.

Kleiber C, Zeileis A (2016). Visualizing count data regressions using rootograms. *The American Statistician* 70(3): 296–303.

Kong M, Xu S, Levy SM, Datta S (2014). GEE type inference for clustered zero-inflated negative binomial regression with application to dental caries. *Computational Statistics and Data Analysis* 85: 54–66.

Lambert D (1992). Zero-inflated poisson regression with an application to defects in manufacturing. *Technometrics 34*(1): 1–14.

Martin TG, Wintle BA, Rhodes JR, Kuhnert PM, Field SA, Low-Choy SJ, Tyre AJ, Possingham HP (2005). Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology Letters* 8(11): 1235–1246.

Miller JM (2007). Comparing poisson, hurdle, and zip model fit under varying degrees of skew and zero-inflation (Doktora Tezi). *University of Florida*.

Misevičiūtė, B. (2001). Elektroninė komercija [PPT]. Retrieved 2018, from http://kopustas.elen.ktu.lt/studentai/lib/exe/fetch.php?media=elektronine_komercija.ppt.

Özen M (2020). Kentsel kavşaklarda trafik kazalarının sıklığını etkileyen faktörlerin incelenmesi. *Teknik Dergi 31*(3): 10033–10053.

Peng J (2013). Count data models for injury data from the national health interview survey (M. Sc. Thesis). *The Ohio State University Graduate Program in Public Health,* Columbus.

Pittman B, Buta E, Krishnan-Sarin S, O'Malley SS, Liss T, Gueorguieva R (2018). Models for analyzing zero-inflated and overdispersed count data: an application to cigarette and marijuana use. *Nicotine&Tobacco Research* 22(8): 1390–1398.

Ridout M, Hinde J, Demetrio CGB (2001). A score test for a zero-inflated poisson regression model against zero ınflated negative binomial alternatives. *Biometrics* 57: 219–233.

Sharma AK, Landge VS (2013). Zero inflated negative binomial for modeling heavy vehicle crash rate on Indian rural highway. *International Journal of Advances in Engineering & Technology* 5(2): 292–301.

Sileshi G (2008). The excess-zero problem in soil animal count data and choice of appropriate models for statistical inference. *Pedobiologia* 52: 1–17.

Tukey JW (1977). Exploratory data analysis. *Addison-Wesley, Reading, MA* 2: 131–160.

Tüzel S (2011). Hasar sıklıkları için sıfır yığılmalı kesikli modeller (Yüksek Lisans Tezi). *Hacettepe Üniversitesi Fen Bilimleri Enstitüsü*, Ankara.

Tüzen MF, Erbaş S (2017). A comparison of count data models with an application to daily cigarette consumption of young persons. *Communications in Statistics Theory And Methods* 47(23): 5825–5844.

Wang Z, Ma S, Wang CY (2015). Variable selection for zero-inflated and overdispersed data with application to health care demand in Germany. *Biometrical Journal* 57(5): 867–884.

Winkelmann R (2000). Econometric analysis of count data (5th edition). *Springer-Verlag Berlin Heidelberg*.

Yang Z, Hardin JW, Addy C (2009). Testing overdispersion in the zero-inflated poisson model. *Journal of Statistics Planning and Inference* 139: 3340–3353.

Yıldırım G (2019). Poisson ve negatif binom regresyon modelleri. (Yüksek Lisans Tezi). *Çukurova Üniversitesi, Fen Bilimleri Enstitüsü*, Adana.

Zhuo L, Stacey K, Lawrence JC, Lisa KH, Richard HLMO (2008). Modeling motor vehicle crashes for street racers using zero-inflated models. *Accident Analysis and Prevention* 40: 835–839.