Tiku'nun Aykırı Değer Belirleme Testi

$x_1, x_2, \ldots, x_n$, $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$ şeklinde tanımlanan herhangi bir dağılımdan rastgele bir örneklem olsun. μ ve σ parametreleri bilinmediğinde, en küçük r_1 veya en büyük r_2 gözlemin aykırı değer olup olmadığını test etmek için Tiku tarafından önerilmiştir.

$$r_1 \geq 0 \ \text{ ve } r_2 \geq 0$$

olmak üzere, Tiku tarafından önerilen test istatistiği

$$T(r_1, r_2) = \frac{\hat{\sigma}_c}{\hat{\sigma}} \qquad (r_1 + r_2 \geq 1)$$

olarak ifade edilir.

Burada,

$\hat{\sigma}_c$: Sansürlenmiş veri için σ nın MML (Uyarlanmış En Çok Olabilirlik) tahmin edicisi

$\hat{\sigma}$: Tam veri için σ nın LS tahmin edicisi

$r_{1,}$ r_2: Önceden belirlenmiş sansürleme sayıları

olarak tanımlanır.

Not: $r_1 = r_2 = 0$ ise $\hat{\sigma}_c = \hat{\sigma}$ olur.

Not: Veri setinde aykırı değer olması $\hat{\sigma}$'yı etkilerken $\hat{\sigma}_c$'yi etkilememektedir.

Normal dağılım için

$$\hat{\sigma}_c = \{B + \sqrt{(B^2 + 4AC)}\}/\{2\sqrt{A(A-1)}\}$$

Burada,

$A = n - r_1 - r_2$

$B = r_2\alpha_2\left(x_{(n-r_2)} - K\right) + r_1\alpha_1\left(x_{(r_1+1)} - K\right)$

$$C = \sum_{i=r_1+1}^{n-r_2} \left(x_{(i)} - K\right)^2 + r_1\beta_1\left(x_{(r_1+1)} - K\right)^2 + r_2\beta_2\left(x_{(n-r_2)} - K\right)^2$$

$$K = \left[\sum_{i=r_1+1}^{n-r_2} x_{(i)} + r_1\beta_1 x_{(r_1+1)} + r_2\beta_2 x_{(n-r_2)}\right] \Big/ m$$

$$m = n - r_1 - r_2 + r_1\beta_1 + r_2\beta_2$$

olarak ifade edilir. Burada,

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

eşitliği kullanılarak hesaplanmaktadır.

Kritik Değer:

$$T_\gamma = \left[\frac{n-1}{n - r_1 - r_2 - 1}u_\gamma + \frac{1}{5n}\left(1 + \frac{1}{n - 2r_2 - 1}\right)\right]\left(\frac{n(A-1)}{A(n-1)}\right)^{1/2}$$

olarak ifade edilir. Burada, u_γ

$$\text{Beta}(n - r_1 - r_2 - 1, r_1 + r_2)$$

dağılımının γ. yüzdelik noktası olarak tanımlanır.

Eğer, $T(r_1, r_2) = \frac{\hat{\sigma}_c}{\hat{\sigma}} < T_\gamma \Rightarrow$ En küçük r_1 ve en büyük r_2 gözlem aykırı değer olarak belirlenir.

<u>Örnek</u>: Aşağıdaki veri setinde, Tiku'nun aykırı değer belirleme testini kullanarak aykırı değer olup olmadığını belirleyelim.

-1.40	-0.05	0.48
-0.44	0.06	0.63
-0.30	0.10	1.01
-0.24	0.18	
-0.22	0.20	
-0.13	0.39	

Normal dağılım varsayımı altında -1.40 ve 1.01'in aykırı değer olup olmadığını belirlemek istediğimizi varsayalım. Bu durumda,

$r_1 = 1$ ve $r_2 = 1$

olarak alınır. Burada, ilgili tablo değerlerine bakıldığında

$n = 15$ için
$\alpha_1 = \alpha_2 = 0.637$
$\beta_1 = \beta_2 = 0.880$

olduğu görünür. Yukarıda verilen eşitlikler yardımıyla,

$$\hat{\sigma} = 0.551$$

ve

$$\hat{\sigma}_c = \left\{0.686 + \sqrt{0.470 + 90.630}\right\}/2\sqrt{156} = 0.410$$

olarak hesaplanır. Bu değerler kullanılarak,

$$T(1,1) = \frac{\hat{\sigma}_c}{\hat{\sigma}} = \frac{0.410}{0.551} = 0.744$$

olarak hesaplanır. Kritik değer,

$$T_{0,10} = \left[\frac{14}{12}(0.732) + \frac{1.0714}{75}\right]\left(\frac{180}{182}\right)^{1/2} = 0.865$$

olarak elde edilir.

$0.744 < 0.865$ olduğundan -1.40 ve 1.01 değerleri aykırı değerdir.

Alternatif Regresyon Metotlarına Giriş

Basit doğrusal regresyon denklemi,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1,2,\dots,n$$

olarak tanımlanır. Burada,
n: Gözlem sayısını
y: Bağımlı değişkeni
x: Bağımsız değişkeni
ε: Hata terimlerini
β_0: Kesim noktasını
β_1: Eğimi
gösterir.

Hata terimlerinin dağılımı normal olduğunda ($\varepsilon_i \sim N(0,\sigma^2)$) β_0 ve β_1 parametrelerinin tahmin edicilerini elde etmek için En Küçük Kareler (Least Squares-LS) yöntemi kullanılır. Çünkü normal dağılım varsayımı altında en etkin tahmin ediciler LS tahmin edicileridir.

β_0 ve β_1 parametrelerinin LS tahmin edicileri

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

olarak tanımlanır.

Ancak söz konusu tahmin ediciler normal dağılım varsayımından sapmalar olduğunda ya da veri setinde aykırı değerler olduğunda etkinliklerini kaybederler. Bir başka deyişle, LS tahmin edicileri normal dağılım varsayımından sapmalara ve aykırı değerlere karşı robust tahmin ediciler değillerdir.

Bu nedenle, literatürde normal dağılım varsayımından sapmalara ve aykırı değerlere karşı robust olan birçok tahmin edici önerilmiştir.

Bir sonraki bölümde, kolay uygulanabilir bazı robust regresyon metotları tanıtılmış ve uygulamaları yapılmıştır.

Veri setinde aykırı değer bulunması LS tahmin yöntemi kullanılarak elde edilen regresyon denkleminin eğimini ciddi oranda etkiler. Bununla beraber, alternatif regresyon yöntemlerine dayalı olarak elde edilen regresyon denkleminin eğimi aykırı değerlerden oldukça sınırlı bir şekilde etkilenir. Bu durum, robust regresyon kullanılarak yapılan çıkarımların daha güvenilir olmasına yol açar.

Burada gösterilen yöntemler daha karmaşık ancak daha etkin robust regresyon yöntemlerini anlamamız için bir temel teşkil etmektedir.