DERLEME REVIEW

DOI: 10.5336/biostatic.2019-71069

Model Performans Kriterlerinin Kronolojisine ve Metodolojik Yönlerine Genel Bir Bakış: Bir Gözden Geçirme

An Overview of Chronology and Methodological Aspects of Model Performance Criteria: A Review

İsmet DOĞAN[a], Nurhan DOĞAN[a]

[a]Afyonkarahisar Sağlık Bilimleri Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim ABD, Afyonkarahisar, TÜRKİYE

ÖZET: Modeller, tahminin en önemli bileşenidir. Ancak her veri kümesi farklı ilişkileri tanımlaması gereken farklı değişken türleri içermektedir ve her model tipinin veri seti ile ilgili kısıtlamaları bulunmaktadır. Bu nedenle, aranan ilişkiyi doğru tanımlayabilecek tahmin modelinin seçilmesi önemlidir. Model seçimi ile ilgili literatürde yer alan çalışmalar, örneklem büyüklüğü, model yapısı, verilerin dağılımı, tahmin yöntemleri ve modelde yer alan değişken sayısı gibi birçok faktörün model seçim kriterlerinin sonuçlarını etkilediğini göstermiştir. Bu durum araştırmacıların en iyi model seçim kriterini ve özelliklerini merak etmelerine neden olmaktadır. Modeli doğrulayan indeksleri kullanmak yerine, modelin verilere uygunluğunu en iyi şekilde değerlendiren uygun indekslerin seçilmesi önerilmesine rağmen, pratikte bu durum oldukça zor ve karmaşıktır. Her ne kadar çalışmalarda bir modeli değerlendirmek için bazı kriterlerin kullanılması önerilmesine rağmen, her bir çalışmada kullanılan veriler birbirinden tamamen farklı olacağından, bu önerilerin genelleştirilemediği görülmektedir. Nicel bir ölçüt olan model değerlendirme kriterleri, tanımlayıcı yeterlilik, basitlik ve genelleştirilebilirlik gibi özellikler içermektedir. Bir modelin yeterliliğini tam olarak değerlendirmek için bu üç özelliğin üçünün de aynı anda değerlendirilmesi gerekmektedir. Çalışmanın amacı, çeşitli performans kriterlerine ve bunların sınıflandırılmasına yönelik genel bir bakış sağlamaktır.

Anahtar kelimeler: Performans ölçüleri; modelleme; model seçimi

ABSTRACT: Models are the most important component of estimation. However, each dataset contains different types of variables that need to define different relationships, and each model type has constraints on the dataset. Thus, it is important to select forecasting model that can define the sought relationship properly. Studies in the literature on model selection have shown that many factors such as sample size, model structure, distribution of data, estimation methods, and number of variables in the model affect the results of the model selection criteria. This makes researchers wonder about the best model selection criteria and features. Although it is recommended to select appropriate indexes that best evaluate the suitability of the model to data rather than using indexes that confirm the model, in practice this is quite difficult and complex. Although some criteria are suggested to evaluate a model in the studies, it is seen that these recommendations cannot be generalized because the data used in each study will be completely different from each other. Model evaluation criteria, which are quantitative measures, include descriptive adequacy (whether the model fits observed data), simplicity (whether the model's description of observed data is achieved in the simplest possible manner) and generalizability (whether the model provides a good predictor of future observations). To fully assess the adequacy of a model, all three of these features need to be evaluated at the same time. The aim of the study is to provide an overview of the various performance criteria and their classification.

Keywords: Performance metrics; modeling; model selection

Matematiksel modeller, farklı alanlardaki süreçleri karakterize etmek için yaygın olarak kullanılmaktadır. Teorik olarak modeller olasılıksal veya istatistiksel hipotez kümeleri olarak değerlendirilmelidir. Matematiksel modellemede amaç, altta yatan sürece en çok yaklaşan modeli tanımlamaktır. Dolayısıyla modelleme

Correspondence: Nurhan DOĞAN
Afyonkarahisar Sağlık Bilimleri Üniversitesi, Tıp Fakültesi Biyoistatistik ve Tıbbi Bilişim ABD, Afyonkarahisar, TÜRKİYE/TURKEY
E-mail: nurhandogan@hotmail.com

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 20 Aug 2019 ***Accepted:*** 14 Nov 2019 ***Available online:*** 02 Dec 2019

2146-8877

ile uğraşan kişilerin modellerin doğrulanması, değerlendirilmesi ve geçerliliği ile ilgili kavramları ayırt etmesi önemlidir. Doğrulama, bir modelin kesinliğinin veya gerçekliğinin / doğruluğunun kanıtıdır. Modelin doğruluğu kısmen verilerin değişkenliğine ve temsil edilebilirliğine bağlıdır. Bu bilgi, modellerin girdi verilerindeki hatalara karşı hassasiyetini değerlendirmek için kullanılabilir. Geçerlilik, ayrıntılı ve bol kanıtlarla modelin biçimsel olarak tanınmasına yol açan bir sonucun oluşturulmasıdır. Değerlendirme ise gözlenen değerler ile tahmin edilen değerlerin karşılaştırıldığı, modele ait bir performans inceleme sürecidir.[1] Modellerin performanslarını incelemek için anlamlı ölçüler ve kriterler gerekmektedir. Ancak, tahmin edebilme yetenekleri değerlendirilen rakip modeller arasında ayrım yapmak, modellerin çeşitli şekillerde farklılık göstermesi (içerdiği değişken sayısı, karmaşıklık vb.) nedeniyle her zaman kolay değildir. Uygun modelin seçilmesi modellemede karşılaşılan temel zorluklardan biridir. Model uyumunun değerlendirilmesinde uyum iyiliği kriterlerinin hesaplanması ve/veya modelde yer alan değişken sayısına bakılması en sık kullanılan yöntemlerdir. Genel olarak uyum iyiliği ile değişken sayısı arasında bir denge bulunmaktadır. Çok değişkenli modeller değişken sayısı az olan modellerden daha iyi bir uyum gösterme eğilimindedir ancak, bu genellikle iyi bir şey değildir. Çünkü daha fazla değişken eklemek genellikle eldeki verilere uygun iyi bir modelle sonuçlanmakta, ancak aynı model diğer veri kümeleri ile ilgili tahminlerde çoğunlukla işe yaramamaktadır. Değişken sayısı ile uyum iyiliği arasındaki doğru dengeyi bulmak ise her zaman mümkün değildir.

Model seçimi ile ilgili karar verme istatistiksel çıkarımda önemli bir sorundur ve sorunun temelinde bir dizi model içerisinden en iyisini seçme düşüncesi yatmaktadır. Model seçim kriterlerinin tümü, hangi modelin en iyi olduğu veya en azından hangisinin en iyi olarak seçilmesi gerektiği sorusunun cevabını sağlayacak şekilde tasarlanmıştır. Modeller onaylanacakları umuduyla değerlendirilirler. Model seçme çalışmalarının ana sorusu "iki rakip model eşdeğer olduğunda, hangisi bilinmeyen gerçek olana daha yakındır" şeklinde ifade edilebilir.[2] Model seçimi nedir? Model seçiminin amaçları nelerdir? Model seçimi yöntemleri nelerdir ve nasıl çalışır? Hangi yöntemler hangi koşullarda diğerlerinden daha iyi performans göstermektedir? Bu sorular, nispeten az gelişmiş bir alandaki bir takım anahtar kavramlara dayanmaktadır. Standart model seçim yöntemleri arasında, sadeliğin (en az varsayıma dayanma ve en az sayıda değişken içerme), uyum iyiliğine karşı dengelendiği bir uygulama sağlayan klasik hipotez testleri, en çok olabilirlik, Bayes yöntemi, minimum tanımlama uzunluğu, çapraz doğrulama ve Akaike'nin bilgi kriteri vb. yer almaktadır. Bu yöntemler örnekleme hatalarını dikkate almaktadırlar ve model seçiminde göreceli başarıları koşullara bağlıdır.[1] En popüler kriterler bile zaman zaman incelenip şiddetle eleştirilmekte ve hatta reddedilmektedir. Hangi kriterin kullanılacağı tartışmaları literatürde yaygındır. Kriterlerin sayısı hızla artma eğilimindedir. Son zamanlarda, yeni kriterler daha düzenli olarak geliştirilmekte ve yayınlanmaktadır. Yeni kriterler geliştirmek için genellikle iki yaklaşım kullanılmaktadır. Birinci yaklaşım, çalışmaya özgü koşullara uyum sağlamak için mevcut kriterleri değiştirmeye, ikinci yaklaşım ise mevcut çeşitli kriterlerde yer alan bilgileri bir araya getirmeye odaklanmaktadır. Yine de, "en iyi kriter" konusunda bir fikir birliği söz konusu değildir. Araştırmacılar, tek bir en iyi kriter için çaba göstermenin gerekmediğine dair daha pratik bir görüş belirtmektedirler. Buna göre "ideal için bir arayış" gerçekçi olmayan bir amaçtır ve tek bir kriterin evrensel olarak en iyisi olması söz konusu değildir. Çünkü kriterler çok sayıda veriyi tek bir değerde yoğunlaştırdığı için, model performansı ile ilgili model hatalarının yalnızca bir yansımasını sağlamaktadır. Performans kriterleri göz önüne alındığında, hatanın rastgele bir değişken olduğu ve bunun tam olarak açıklanmasının ancak varsa olasılık yoğunluk fonksiyonu veya momentlerle mümkün olacağı, belirtilmesi gereken temel noktalardan biridir.[3] Model seçim kriterleri hiçbir zaman bir hipotezin doğruluğu veya yanlışlığı ile ilgili değil, hipotezin gerçeğe yakınlığı ile ilgilidir. Model seçiminde yanlılık ve varyans arasındaki denge dikkate alınmaktadır. Model seçimi, belirlenmiş bir kayıp fonksiyonu ile belirli bir amaç için modellerden birinin seçiminin hedeflendiği bir model seti ile başlar. Model seçimi, incelenen tüm modelleri simetrik olarak ele alır, ancak

hipotez testi asimetrik olarak yokluk ve alternatif hipoteze farklı bir statü yükler. Hipotez testinde, yokluk hipotezinden bir veya daha fazla alternatif hipotez yönünde ayrılmanın istatistiksel olarak anlamlı bir kanıtın olup olmadığı sorgulanmaktadır. Model seçimi için hipotez testi kullanılması durumunda söz konusu olan tüm modellerin reddedilmesi mümkündür. Klasik hipotez testleri için yokluk hipotezi farklı rollere sahiptir.[1] Model seçimi ile ilgili karar verme aşamasında araştırmacılar için performans kriterleri kritik öneme sahip hayati unsurlardır. Dolayısıyla makalenin pratikteki önemi, performans kriterlerini kullanıcıların öğrenmesini kolaylaştırmak, uygulayıcılar için kriter seçim sürecini hızlandırmak ve akademisyenler için yeni kriter geliştirme çalışmaları için kullanılabileceği gerçeğidir. Çalışmanın amacı, çeşitli performans kriterlerine ve bunların sınıflandırılmasına yönelik genel bir bakış sağlamaktır. Bu çalışmanın ana katkısı, performans kriterleri ile ilgili matematiksel formül ve görselleştirme çizelgesi yardımıyla özelliklerinin anlaşılmasını sağlamak, önceki çalışmalardan yararlanarak modeller arasında seçim yapmak için önerilerde bulunmaktır. Makalede Helsinki Deklerasyonu Prensipleri dikkate alınmıştır.

GEREÇ VE YÖNTEMLER

Modern bilimsel çalışmalar, model seçimi sorunuyla ilgilenmektedir. Araştırmacılar genellikle, gözlemlenen birimlerin farklı yönleriyle ilgili ölçümler şeklinde veri toplamakta ve bu değişkenlerin bazı ilgi alanlarını nasıl etkilediğini incelemek istemektedirler. Dikkate alınan değişkenler arasında etkileşim var mı? Bunların hangileri sonuçlar üzerinde etkilidir, hangileri etkili değildir? vb. sorular araştırmacılar tarafından cevaplanması gereken sorulardır. Dolayısıyla yıllar içinde model seçimi gibi önemli bir soruna çözüm bulmak için birçok kriter önerilmiş olması şaşırtıcı değildir.[4] Ağırlıklı olarak 1960'lı yıllardan itibaren, model çıktısının doğruluğunu değerlendirmek için bir takım uygunluk kriterleri önerilmiştir. Bu çalışmada, literatürde yer alan kriterlerin bazı temel kavramlar dikkate alınarak seçici bir araştırması yapılmış ve belirlenen kriterler, kavramsal bir bakışla Tablo1-6'da görüldüğü üzere altı farklı başlık altında sınıflandırılmışlardır.

TABLO 1: Korelasyon katsayısına dayalı kriterler.

Kriter	Tanım	Kronoloji
r	$\frac{n^{-1}\Sigma_{i=1}^{n}(O_i-\overline{O})(P_i-\overline{P})}{s_P s_O}$	*Pearson, 1948*[5] *Fox., 1981*[6] *Muroi et al., 2015*[7] *Duveiller et al., 2016*[8]
R^2	$\left\{\frac{\Sigma_{i=1}^{n}(O_i-\overline{O})(P_i-\overline{P})}{\left[\Sigma_{i=1}^{n}(O_i-\overline{O})^2\right]^{0.5}\left[\Sigma_{i=1}^{n}(Pi-\overline{P})^2\right]^{0.5}}\right\}^2$	*Theil, 1961*[9] *Legates and McCabe,* 1999[10] *Krause et al,, 2005*[11]
$R_{Adjusted}$	$1-\frac{(n\text{-}1)(1\text{-}R^2)}{n\text{-}p}$	*Theil, 1961*[9]
R_{KG}	$1-\sqrt{(r\text{-}1)^2+\left(\frac{s_O}{s_P}\text{-}1\right)^2+\left(\frac{O}{P}\text{-}1\right)^2}$	*Gubta et al., 2009*[12] *Kling et al., 2012*[13]

P_i:i=1.2, ..., n→Tahmin edilen değerler, Oi:i=1.2, ..., n→Gözlenen değerler, n: Gözlem sayısı, $\overline{O}$: Gözlenen değerlerin (O_i) ortalaması, $\overline{P}$: Tahmin edilen değerlerin (P_i) ortalaması, s_p ve s_O sırasıyla P_i ve O_i değerlerinin standart sapmaları, p: Değişken sayısı, r: Pearson korelasyon katsayısı, R^2 : Belirtme katsayısı, $R^2_{Adjusted}$: Düzeltilmiş belirtme katsayısı, R_{KG} : Kling-Gubta belirtme katsayısı.

TABLO 2: Bilgi ölçütlerine dayalı kriterler.

Kriter	Tanım	Kronoloji
AIC	$n * In\left(\frac{SSE}{n}\right) + 2 * p$	*Akaike, 1974*[14]
BIC_{Sawa}	$n * In\left(\frac{SSE}{n}\right) + \frac{2 * (p+2) * n * s^2}{SSE} - \frac{2 * n^2 * s^4}{SSE^2}$	*Sawa, 1978*[15]
$BIC_{Schwarz}$	$n * In\left(\frac{SSE}{n}\right) + p * In(n)$	*Schwarz, 1978*[16]
HQ	$n * In\left(\frac{SSE}{n}\right) + p * In(In(n))$	*Hannan and Quinn, 1979*[17]
AICC	$n * In\left(\frac{SSE}{n}\right) + \frac{2 * (p+1)}{n\text{-}p\text{-}2}$	*Hurvich and Tsai, 1989*[18]

n: Gözlem sayısı, p: Değişken sayısı, s^2: Tam modelden elde edilen hata varyansı, SSE: Error sum of squares, AIC: Akaike Information Criterion, AICC: Corrected Akaike Information Criterion, $BIC_{Schwarz}$: Schwarz Bayesian Information Criterion, HQ: Hannan-Quinn Criterion, BIC_{Sawa}: Sawa Bayesian Information Criterion.

TABLO 3: Tahmin edilen varyans değerine dayalı kriterler.

Kriter	Tanım	Kronoloji
URV	$\hat{\sigma}_p^2 + \frac{p * \hat{\sigma}_p^2}{n * (1 - p/n)}$	*Theil, 1961*[9]
C_p	$\hat{\sigma}_p^2 + 2 * p * \hat{\sigma}_p^2 / n$	*Mallows, 1973*[19]
S_p	$\hat{\sigma}_p^2 + \frac{p * \hat{\sigma}_p^2 * (2 - (p + 1/n))}{[n * (1 - p/n)\ (1 - (p+1)/n)]}$	*Hocking, 1976*[20]
GCV	$\hat{\sigma}_p^2 + \frac{p * \hat{\sigma}_p^2 * (2 - p/n)}{n * (1 - p/n)^2}$	*Golub et al., 1979*[21]
PC	$\hat{\sigma}_p^2 + \frac{2 * p * \hat{\sigma}_p^2}{n * (1 - p/n)}$	*Amemiya, 1980*[22]
BEC	$\hat{\sigma}_p^2 + p * \hat{\sigma}_p^2 * In\ (n) / n$	*Geweke and Meese, 1981*[23]
S	$\hat{\sigma}_p^2 + \frac{2 * p * \hat{\sigma}_p^2}{n}$	*Shibata, 1981*[24]
MDL	$In\ (n * \hat{\sigma}_p^2) + \frac{In(\lvert X'X \rvert)}{n}$	*Rissanen, 1987*[25]

σ_p^2: σ^2 'ye ait, dikkate alınan modelden elde edilen en çok olabilirlik tahmini, n: Gözlem sayısı, p: Değişken sayısı, $\lvert X'X \rvert$: $X'X$ matrisinin determinantı, URV: Unbiased Residual Variance, PC: Prediction Criterion, BEC: Bayesian Estimation Criterion, GCV: Generalized Cross-Validation, MDL: Minimum Description Length.

TABLO 4: Hemfikir olma/etkinlik kriterleri.

Kriter	Tanım	Kronoloji
A	$1-\frac{\Sigma_{i=1}^{n}(O_i-Z_i)^2+\Sigma_{i=1}^{n}(P_i-Z_i)^2}{\Sigma_{i=1}^{n}(O_i-\overline{Z})^2+\Sigma_{i=1}^{n}(P_i-\overline{Z})^2}$	*Robinson, 1957*[26]
E	$1-\frac{\Sigma_{i=1}^{n}(P_i-O_i)^2}{\Sigma_{i=1}^{n}(O_i-\overline{O})^2}$	*Nash and Sutcliffe, 1970*[27]
d	$1-\frac{\Sigma_{i=1}^{n}(O_i-P_i)^2}{\Sigma_{i=1}^{n}(\lvert P_i-\overline{O}\rvert+\lvert O_i-\overline{O}\rvert)^2}$	*Willmott, 1981*[28]
ρ	$1-\frac{MAE}{\frac{1}{n^2}\Sigma_{i=1}^{n}\Sigma_{i=1}^{n}\lvert P_j-O_i\rvert}$	*Mielke, 1991*[29]
E_{LG}	$\frac{\Sigma_{i=1}^{n}(O_i-\overline{O})^2+\Sigma_{i=1}^{n}(O_i-P_i)^2}{\Sigma_{i=1}^{n}(O_i-\overline{O})^2}$	*Loague and Green, 1991*[30]
M	$(2/\pi)sin^{-1}\left\{1-\frac{MSE}{s_p^2+s_o^2+(\overline{P}-\overline{O})^2}\right\}$	*Watterson, 1996*[31]
E_1	$1-\frac{\Sigma_{i=1}^{n}\lvert P_i-O_i\rvert}{\Sigma_{i=1}^{n}\lvert O_i-\overline{O}\rvert}$	*Legates and McCabe, 1999*[10]
$d_{Relative}$	$1-\frac{\Sigma_{i=1}^{n}\left(\frac{O_i-P_i}{O_i}\right)^2}{\Sigma_{i=1}^{n}\left(\frac{\lvert P_i-O\rvert+\lvert O_i-O\rvert}{\overline{O}}\right)^2}$	*Krause et al., 2005*[11]
$E_{Relative}$	$1-\frac{\Sigma_{i=1}^{n}\left(\frac{O_i-P_i}{O_i}\right)^2}{\Sigma_{i=1}^{n}\left(\frac{O_i-\overline{O}}{\overline{O}}\right)^2}$	*Krause et al., 2005*[11]
AC	$1-\frac{\Sigma_{i=1}^{n}(O_i-P_i)^2}{\Sigma_{i=1}^{n}(\lvert \overline{O}-\overline{P}\rvert+\lvert O_i-\overline{O}\rvert)(\lvert \overline{O}-\overline{P}\rvert+\lvert P_i-\overline{P}\rvert)}$	*Ji and Gallo, 2006*[32]

TABLO 4: Hemfikir olma/etkinlik kriterleri (devam).

(Devam).	Tanım	Kronoloji
FIT_{Ratio}	$\left\{1-\frac{\sqrt{\sum_{i=1}^{n}(O_i-P_i)^2}}{\sqrt{\sum_{i=1}^{n}(O_i-\overline{O})^2}}\right\}*100$	*Muroi et al., 2015*[7]

P_i:i=1.2, ..., n→Tahmin edilen değerler, O_i:i=1.2, ..., n→Gözlenen değerler, n: Gözlem sayısı, $\overline{O}$: Gözlenen değerlerin (O_i) ortalaması, $\overline{P}$: Tahmin edilen değerlerin (P_i) ortalaması, s_p ve s_o sırasıyla P_i ve O_i değerlerinin standart sapmaları, p: Değişken sayısı, Z_i: O_i ve P_i *değerlerinin ortalaması*, $\overline{Z}$: $\overline{O}$ ve $\overline{P}$ değerlerinin ortalaması, MAE: Mean Absolute Error, MSE: Mean Square Error, AC: Agreement Coefficient.

Önemli Not:

$FIT_{Ratio}=\left\{1-\frac{\sqrt{\sum_{i=1}^{n}(O_i-P_i)^2}}{\sqrt{\sum_{i=1}^{n}(O_i-\overline{O})^2}}\right\}*100$ ifadesinde yer alan $\frac{\sqrt{\sum_{i=1}^{n}(O_i-P_i)^2}}{\sqrt{\sum_{i=1}^{n}(O_i-\overline{O})^2}}$ ifadesi literatürde farklı bir kriter olarak RSR (RMSE-observations standard deviation ratio) ismi ile ifade edilmektedir.

TABLO 5: Göreli hataya dayalı kriterler.

Kriter	Tanım	Kronoloji
SE	$\left[\frac{1}{n}\sum_{i=1}^{n}\left\|O_i-P_i\right\|^2\right]^{1/2}$	*Fisher, 1920*[33]
MSE	$\frac{1}{n}\sum_{i=1}^{n}(O_i-P_i)^2$	*Fisher, 1920*[33]
RMSE	$\left[\frac{1}{n}\sum_{i=1}^{n}(O_i-P_i)^2\right]^{0.5}$	*Patry and Marino, 1983*[34]
RMSPE	$\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{P_i-O_i}{P_i}*100\right)^2\right]^{0.5}$	*Manley, 1978*[35]
RRMS	$\frac{RMSE}{\overline{O}}*100$	*Loague and Green,1991*[30]
SAE, *MAE*	$\sum_{i=1}^{n}\left\|P_i-O_i\right\|, \frac{1}{n}\sum_{i=1}^{n}\left\|P_i-O_i\right\|$	*Stephenson, 1979*[36] *Legates and McCabe, 1999*[10]
SRMSE	$\frac{1}{\overline{O}}*\sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i-O_i)^2}$	*Dust et al., 2000*[37]
MAPE, AARE	$\frac{1}{n}\sum_{i=1}^{n}\frac{\left\|P_i-O_i\right\|}{P_i}*100$	*Nayak et al., 2004*[38]
MBE	$\frac{1}{n}\sum_{i=1}^{n}(P_i-O_i)$	*Ji and Gallo, 2006*[32]
MAD	$\frac{1}{n}\sum_{i=1}^{n}\left\|O_i-\overline{O}\right\|$	*Willmott et al., 2012*[39]

P_i:i=1.2, ..., n→Tahmin edilen değerler, O_i:i=1.2, ..., n→Gözlenen Değerler, n: Gözlem sayısı, $\overline{O}$: Gözlenen değerlerin (O_i) ortalaması, p: Değişken sayısı, SE: Standard Error, MSE: Mean Square Error, RMSE: Root Mean Square Error, RMSPE: Root Mean Square Percentage Error, SAE: Sum of Absolute Error MAE: Mean Absolute Error, MAD: Mean Absolute Deviation, RRMS: Relative Root Mean Square, SRMSE: Scaled Root Mean Square Error, MBE: Mean Bias Error, MAPE: Mean Absolute Percentage Error, AARE: Average Absolute Relative Error,.

TABLO 6: Kovaryansa dayalı kriterler.

Kriter	Tanım	Kronoloji
$\chi 2$	$tr(E^{-1}S-I)-log\lvert E^{-1}S\rvert$	*Joreskog,1973*[40]
NNFI	$\frac{\chi^2_{null}/df_{null}-\chi^2_{model}/df_{model}}{(\chi^2_{null}/df_{null})-1}$	*Tucker and Lewis, 1973*[41]
Normed $\chi 2$	χ^2/df	*Wheaton et al., 1977*[42]
NFI	$\frac{(\chi^2_{null}-\chi^2_{model})}{\chi^2_{null}}$	*Bentler and Bonett, 1980*[43]
GFI	$1-\frac{\chi^2_{model}}{\chi^2_{null}}$	*Joreskog and Sorbom, 1981*[44]
AGFI	$1-\frac{\chi^2_{model}/df_{model}}{\chi^2_{null}/df_{null}}$	*Joreskog and Sorbom, 1981*[44]
RMR	$\sqrt{\frac{2*\Sigma_{i=1}^{p}\Sigma_{j=1}^{i}(s_{ij}-e_{ij})^2}{p*(p+1)}}$	*Joreskog and Sorbom, 1981*[44]
SRMR	$\sqrt{\frac{2*\Sigma_{i=1}^{p}\Sigma_{j=1}^{i}[(s_{ij}-e_{ij})/(s_{it}s_{jj}2)]^2}{p*(p+1)}}$	*Joreskog and Sorbom, 1981*[44]
CN	$\left\{\frac{(z_{crit}+\sqrt{2*df-1})^2}{2*\chi^2/(n-1)}\right\}+1$	*Hoelter, 1983*[45]
IFI	$\frac{\chi^2_{null}-\chi^2_{model}}{\chi^2_{null}-df_{model}}$	*Bollen, 1989*[46]
CFI	$1-\frac{max[(\chi^2_{model}-df_{model}),0]}{max[(\chi^2_{model}-df_{model}),(\chi^2_{null}-df_{null}),0]}$	*Bentler, 1990*[47]
NI	$\frac{(\chi^2_{null}-df_{null})-(\chi^2_{model}-df_{model})}{\chi^2_{null}-df_{null}}$	*McDonald and Marsh, 1990*[48]
ECVI	$\frac{\chi^2}{n-1}+\frac{2*p}{n-1}$	*Browne and Cudeck, 1992*[49]
RMSEA	$\sqrt{\frac{(\chi^2-df)}{df*(n-1)}}$	*Steiger, 2000*[50]

E: Türetilmiş kovaryans matrisi, S: Gözlenen değerlere ait kovaryans matrisi, n: Gözlem sayısı, s_{ii}, s_{jj}, s_{ij} ve e_{ij} S ve E matrislerine ait değerler, df=0.5 x p x (p+1)−t: Serbestlik derecesi, p: Değişken sayısı, t: Tahmin edilen bağımsız parametre sayısı, χ^2_{null}: Esas alınan (baseline) modele ait Ki-kare değeri, χ^2_{model}: Test edilen modele ait Ki-kare değeri, z_{crit}: Belirli bir olasılık değeri için normal dağılım tablosuna ait kritik değer, *ECVI*: Expected Cross-Validation Index, GFI: Goodness of Fit Index, AGFI: Adjusted Goodness of Fit Index, CFI: Comparative Fit Index, NFI: Normed Fit Index, NNFI: Non-Normed Fit Index, IFI: Incremental Fit Index, NI: Noncentrality Index, CN: Critical N, RMR: Root Mean Residual, SRMR: Standardized Root Mean Residual, RMSEA: Root Mean Squared Error of Approximation.

Tipik istatistiksel çıkarım prosedürleri parametrik ya da parametrik olmayan modelleri belirleyerek veriler ile ilgili tahminlerde bulunur. Bununla birlikte, herhangi bir veri ve amaç için genel olarak uygun bir model yoktur. Uygun olmayan bir model veya yöntem seçimi tamamen ciddi yanıltıcı sonuçlara veya hayal kırıklığı

yaratan tahmin performanslarına neden olabilir. Bu nedenle, tipik bir veri analizinde çok önemli bir adım, bir dizi aday modeli göz önünde bulundurmak ve sonra en uygun olanı seçmektir. Başka bir deyişle model seçimi, veri seti verilen bir dizi model içerisinden istatistiksel bir model seçme işidir.[51] Araştırmacıların bir modelin performansını değerlendirmek zorunda olmalarının nedenleri;

- Modelin geçmişteki ve/veya gelecekteki davranışını yeniden üretme yeteneğinin niceliksel bir tahminini sağlamak,

- Modele ait parametre değerlerinin düzeltilmesi için bir araç sağlamak,

- Mevcut modelleme çabalarını önceki çalışma sonuçları ile karşılaştırmak,

olarak sıralanabilir.[11] Model performansının değerlendirilmesi, yani model tarafından üretilen tahminlerin gözlenen değerlerle karşılaştırılması, model geliştirme ve kullanım için temel bir adımdır. Modeller elde edildiğinde, bazı yönlerini doğrulamak gerekir. Bu doğrulama süreci genellikle, modelin tahminlerinin ne kadar iyi üretildiğinin, gözlemlenen değerleri nasıl simüle ettiğini gösteren matematiksel ölçümlere dayanan bir kriter tanımı içerir.[52] Model performans kriterleri genel olarak, gözlemlerdeki değişkenliğin bir ölçüsü ile normalleştirilen hata teriminin (gözlenen değer ile tahmin edilen değer arasındaki fark) toplamını içermektedir. Performans kriterlerinin birçoğu karşıt işaretlere sahip hataların iptal edilmesini önlemek için, hatalara ait mutlak değerleri veya hataların karelerinin toplamını kullanmaktadır. Sonuç olarak, daha büyük hatalara önem verilirken, daha küçük hatalar ihmal edilme eğilimindedir.[11] Seçilen modelin uyumlu olması durumunda, araştırmacılar tarafından;

- İstatistiğin seçilen modelden anlamlı sapmalara karşı gücü değerlendirilmelidir. Çünkü istatistiğin seçilen model ile önemli ölçüde anlamlı alternatif modeller arasında ayrım yapma gücünün olmaması iyi olabilir.

- Model genel olarak uygun olmasına rağmen, bazı verilerin model tarafından iyi tahmin edilip edilmediğini incelemek için seçilen modelin parça parça bir değerlendirmesi yapılmalıdır.

- Seçilen modelden deneysel olarak ayırt edilemeyen modellerin var olup olmadığı dikkate alınmalıdır.

Modelin uyumlu olmaması durumunda ise, araştırmacılar tarafından;

- Uyumsuzluğun kaynağını belirlemek için modelin parçalı bir değerlendirmesi yapılmalıdır.

- Gözlenen değerler ile tahmin edilen değerler arasındaki tutarsızlık değerlendirilmeli, standartlaştırılmış her bir tutarsızlığın büyüklüğü denetlenmelidir.

Model uyumlu olsun ya da olmasın gerçekte araştırmacılar, her durumda rakip alternatifler arasından bir model seçerken, modelde yer alacak parametre sayısının azlığı için çaba göstermelidirler. Çok sayıda değişken modellendiğinde, modelin iyi bir şekilde uyum göstermesini beklemek gerçekçi değildir. Çünkü verilerin bazı bölümleri model tarafından iyi bir şekilde üretilmeyeceği için genel test istatistiklerinin değeri büyük olacaktır.[53] Literatürde birçok model seçim kriteri önerilmektedir. Araştırmacıların duyarlılık (değişkenler arasındaki ilişkileri yeterince modellemek için yeterli parametreye sahip olmak) ile spesifikliği (bir modele uymamak veya var olmayan ilişkileri öne sürmemek) dengelemek zorunda oldukları unutulmamalıdır.[54]

TARTIŞMA VE SONUÇ

Model spesifikasyonu, modellenen fenomen hakkında tam bilgi sahibi olunması her zaman mümkün olmadığı için zordur. Ampirik veriler çoğunlukla, dikkate alınan fenomeni etkileyen değişkenler ile fenomenin özellikleri hakkında kısmi bilgi temin ettiğinden sınırlıdır. Sınırlı bilgiyle, gerçek modeli oluşturmak ise imkansızdır. Üstelik sınırlı bilgilerle, birden fazla modelin makul olması da muhtemeldir. Model seçimi ile ilgili kriter geliştirme konusu, matematiksel istatistikte diğer konulara göre nispeten yeni bir konudur. Üstelik bu alandaki temel kavramlar karmaşık ve anlaşılması zor olan kavramlardır. Standart model seçim yöntemleri, model

parametrelerinin tahminindeki hataları telafi edebilir. Parametre sayısı arttıkça, telafi değeri artar, bu da eğer bir modelin sadeliği parametre sayısının azlığı ile ölçülürse basitlikle uyum sağlamaları anlamına gelir.[2] Bir modelin verilere yeteri kadar uyum sağlaması, modelin dikkate alınmaya değer aday modeller listesinde yer almasını sağlar. Ancak uyum iyiliği kriterlerine göre iyi olan ve verilere iyi uyum gösteren birkaç farklı model olması mümkündür. Bu durum model seçiminde bir sorundur. Dolayısıyla, farklı birkaç model arasından seçim nasıl yapılmalıdır? Model seçiminde amaç, altta yatan sürece en yakın yaklaşımı temsil eden bir dizi aday model arasından birini seçmektir. Bir veri seti için seçilen en uygun model her zaman hedeflenen amacı gerçekleştiremeyebilir. Bunun nedeni bir modelin, modelin doğruluğu ile ilgili olmayan nedenlerden dolayı da rakiplerine göre üstün bir uyum sağlayabilmesidir. Örneğin, çoğunlukla çok parametreli ve doğrusal olmayan karmaşık bir model, birkaç parametreli basit bir modele göre veriye daha iyi uyum sağlayacaktır. Bu durum, aşırı uyum olarak ifade edilmektedir ve aşırı uyumdan kaçınmak, her model seçim kriterinin başarmak istediği şeydir. Modern model seçim kriterlerinin arkasındaki temel fikir, verilerin içsel olarak gürültülü olması nedeniyle ideal bir modelin gürültüyü değil, sadece temel olguyu yakalayan model olduğunun farkına varmaktır. Gürültü, belirli bir veri kümesine özel olmadığı için, gürültü içeren bir model gelecekteki olaylar hakkında kötü tahminlerde bulunacaktır. Bu, model seçiminde günümüzün "altın standardı" na, "genelleştirilebilirliğe" yol açar. Genelleştirilebilirlik, bir modelin gözlenen veri kümesi ile aynı süreçten gelen verilere ait istatistikleri tahmin etme kabiliyetini ifade eder.[55] Birçok alanda çok çeşitli kriterler önerilmiş ve kullanılmıştır. SRMR, RMSEA, CFI ve TLI uygunluk kriterleri gerek örneklem büyüklüğüne ve değişkenlerin normal dağılım varsayımına karşı duyarsız oldukları için gerekse AIC, ECVI, NFI, BFI, ve AGFI kriterlerine göre model tespitinde daha hassas oldukları için sıklıkla önerilmektedirler.[56] Willmott ve Matsuura (2005) RMSE'nin, model performansı için iyi bir kriter olmadığını ve ortalama hatanın yanıltıcı bir göstergesi olabileceğini ve dolayısıyla MAE'nin bu amaç için daha iyi bir kriter olduğunu ileri sürmüşlerdir. Son 25 yılda gerçekleştirilen birçok çalışmada MSE, RMSE, MAE ve MAPE en popüler kriterler olarak tespit edilmiştir.[3,57] Ancak hata kareler toplamına dayalı kriterler genellikle ortalama sapma, ortalama hata ve ortalama değişkenlik gibi belirsiz göstergeler olarak kabul edilir. Bu kriterlerin çok az bilinen ve istenmeyen özellikleri, sık sık yanlış kullanılmalarına ve yanlış yorumlanmalarına sebep olur. Kareler toplamına dayalı bir kriterin ortalama hatayı, ortalama sapmayı veya ortalama değişkenliği güvenilir bir şekilde temsil edebileceği üstü kapalı şekilde varsayıldığı için tipik olarak zorluklar ortaya çıkar. Bu kriterlerin değerlerinin açık bir şekilde kesin bir yorumu yoktur, çünkü kareler toplamına dayalı kriterler, hem merkezi eğilim hem de hata değişkenliği ve sapma büyüklüğüne göre değişmektedir.[58] AIC, BIC ve AICC kriterleri karşılaştırılmış, doğru modelin seçiminde basit bir model kullanıldığında, BIC ve AICC kriterlerinin AIC kriterinden, karmaşık bir model kullanıldığında ise AIC kriterinin BIC ve AICC kriterlerinden daha yüksek seçim olasılıklarına sahip olduğu ifade edilmiştir.[59,60] Uyumu veya uyumsuzluğu değerlendirmek için r, $R2$, MAE, RMSE vb. birçok istatistiksel yöntem yaygın olarak kullanılmaktadır. Ancak, bu geleneksel kriterler, uyum veya uyumsuzluğu değerlendirmek için her zaman uygun değildir. Örneğin, r veya $R2$ yaygın olarak model değerlendirmesi için kullanılmasına rağmen, bu istatistikler aykırı değerlere karşı aşırı duyarlı, gözlenen değerler ile tahmin edilen değerler arasındaki farklara ise duyarsızdır. Üstelik bu değerler iki veri seti arasındaki gerçek farktan ziyade sadece bunlar arasında doğrusal kovaryasyon olup olmadığını göstermektedir. $R2$, kovaryans ile gözlenen ve tahmin edilen değerlerin çarpılmış standart sapmaları arasındaki karesel oran olarak da ifade edilmektedir. Bu nedenle, gözlenen ve tahmin edilen değişkenlerin ayrı ayrı dağılımlarına karşı birleşik dağılımı tahmin etmektedir. Ayrıca $R2$, gözlenen değerlere ait dağılımın ne kadarının tahmin edilen değerlere ait dağılım ile açıklandığını göstermektedir. Dolayısıyla dağılımın sadece $R2$ ile nicelleştirilmesi $R2$'nin en büyük dezavantajlarından biridir. Çünkü tahmin edilen değerlerin tümü yanlış olsa dahi sistematik olarak yüksek veya düşük tahminlerde bulunan bir model, 1'e yakın $R2$ değerlerine sahip bir model olabilir. MAE ve RMSE ise boyutsal uyumsuzluk kriterleri olduğundan değişkenin ölçeğinden ve biriminden bağımsız değildir. Willmott (1981, 1982) tarafından geliştirilen uyum kriteri, yukarıda belirtilen kriterlerin bazı dezavantajlarını gidermektedir. Dolayısıyla, Wil-

lmott tarafından ortaya konulan kriter modellerin değerlendirilmesi için daha uygundur.[11,32] Çok değişkenli modellerin değerlendirilmesinde kullanılan uyum iyiliği indekslerinin matematiksel özelliklerini açıklamak için indeksler merkezsizlik (noncentrality) parametresinin işlevleri olarak ifade edilmektedir. Merkezsizlik parametresinin normlandırılmış fonksiyonu, yansız mutlak uyum indeksi olarak Tucker-Lewis indeksi ile Bentler-Bonett indeksinin alternatifi olarak önerilmektedir. İncelenen indekslerin çoğunun sistematik olarak örneklem büyüklüğünden etkilendiği gösterilmiştir. Akaike bilgi kriterinin, gerçek uygulamalarda model seçiminde kullanılamayacağı ileri sürülmektedir.[48,61] Spesifik olarak hidrolojik modellerin değerlendirilmesinde MSE ve Nash-Sutcliffe verimlilik kriterleri en yaygın kullanılan iki kriterdir. MSE'nin değeri tahmin edilen değişkenin birimine bağlıdır. Nash-Sutcliffe verimlilik kriteri ise boyutsuzdur. Dolayısıyla, Nash-Sutcliffe verimlilik kriteri genel olarak model performansının raporlanması (ve karşılaştırılması) için tercih edilen bir ölçüdür.[12] Kling-Gupta verimlilik kriteri ise, Nash-Sutcliffe verimlilik kriterinin ayrışmasına dayanmaktadır.[62]

Modeller karar vermede yardımcı olmak için gerekli ve istenen araçlardır ve modellerin belirli bir zamanda ve yerde elde edilen verilerin tahmini değerini temsil eden bir sonuç ürettiği kabul edilmelidir. Karar vericiye yalnızca tahmini bir değer vermekle kalmayıp, aynı zamanda bu değerle ilgili hata miktarının bir tahmininin elde edilmesini sağlamasıda önemlidir. Hem tahmin edilen değer hem de tahmin edilen hata miktarı, gerçek değerler için bir güven aralığı sağlamak üzere birleştirilebilir. Verileri yeterince iyi tanımlayan, en iyi tahminde bulunan, en basit model en uygun modeldir. Dolayısıyla, bilimsel teorinin önemli amaçlarından biri de doğru tahminler üreten modelleri belirlemeyi sağlayacak kriterler ortaya koymak olmalıdır. Farklı kaynaklardan elde edilen veri kümeleri arasındaki karşılaştırmalarda tüm veri kümelerinin bazı ölçüm hataları içerdiği varsayılır. Her zaman küçük de olsa ölçüm hatası vardır ve veri toplama işlemi sırasında bu hatayı arttıran başka kontrolsüz değişkenlik kaynakları da olabilir. Hata verideki düzenliliği bulanıklaştırmakta, modelleme zorluğunu artırmaktadır. Dolayısıyla, bu tür veri kümeleri arasındaki karşılaştırmalar için geçerli bir uyum kriteri geliştirmek gereklidir. Seçim kriterlerinin, araştırmanın sonuna değil yalnızca araştırmaya rehberlik ettiği unutulmamalıdır. Birçok yönden, model seçim kriterlerini model karşılaştırma veya değerlendirme aracı olarak düşünmek daha iyi olacaktır, çünkü seçim terimi daha kesin bir şeyin elde edildiği düşüncesini içermektedir. Model seçiminde kullanılan kriterler gerekliliğin ötesinde çoğaltılmamalıdır.

Finansal Kaynak

Bu çalışma sırasında, yapılan araştırma konusu ile ilgili doğrudan bağlantısı bulunan herhangi bir ilaç firmasından, tıbbi alet, gereç ve malzeme sağlayan ve/veya üreten bir firma veya herhangi bir ticari firmadan, çalışmanın değerlendirme sürecinde, çalışma ile ilgili verilecek kararı olumsuz etkileyebilecek maddi ve/veya manevi herhangi bir destek alınmamıştır.

Çıkar Çatışması

Bu çalışma ile ilgili olarak yazarların ve/veya aile bireylerinin çıkar çatışması potansiyeli olabilecek bilimsel ve tıbbi komite üyeliği veya üyeleri ile ilişkisi, danışmanlık, bilirkişilik, herhangi bir firmada çalışma durumu, hissedarlık ve benzer durumları yoktur.

Yazar Katkıları

All authors contributed equally while this study preparing.

KAYNAKLAR

1. Forster MR. Key concepts in model selection: performance and generalizability. J Math Psychol. 2000;44(1):205-31. [Crossref] [PubMed]
2. Sayyareh A, Obeidi R, Bar-Hen A. Empiricial comparison between some model selection criteria. Commun Stat Simulat. 2010;40(1):72-86. [Crossref]
3. Botchkarev A. A new typology design of performance metrics to measure errors in machine learning regression algorithms. IJIKM. 2019;14:45-76. [Crossref]
4. Kadane JB, Lazar NA. Methods and criteria for model selection. J Am Stat Assoc. 2004;99(465):279-90. [Crossref]
5. Pearson K. Karl Pearson's Early Statistical Papers. 1st ed. Cambridge: Cambridge University Press; 1948. p.557.
6. Fox DG. Judging air quality model performance. B Am Meteorol Soc. 1981;62(5):599-609. [Crossref]

7. Muroi H, Takeshita Y, Adachi S. Model validation criteria for system identification in time domain. T Soc Instr Control Eng. 2015;51(10):80-91. [Crossref]
8. Duveiller G, Fasbender D, Meroni M. Revisiting the concept of a symmetric index of agreement for continuous datasets. Sci Rep-UK. 2016;6:19401. [Crossref] [PubMed] [PMC]
9. Theil H. Economic Forecasts and Policy. 2nd ed. Amsterdam: North-Holland Pub Co;1961. p.657.
10. Legates DR, McCabe GJ. Evaluating the use of goodness of fit measures in hydrologic and hydroclimatic model validation. Water Resour Res. 1999;35(1):233-41. [Crossref]
11. Krause P, Boyle DP, Base F. Comparison of different efficiency criteria for hydrological model assessment. ADGEO. 2005;5:89-97. [Crossref]
12. Gupta HV, Kling H, Yilmaz KK, Martinez GF. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J Hydrol. 2009;377(1-2):80-91. [Crossref]
13. Kling H, Fuchs M, Paulin M. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. J Hydrol. 2012;424-425:264-77. [Crossref]
14. Akaike H. A new look at the statistical model identification. IEEE T Automat Contr. 1974;19(6):716-23. [Crossref]
15. Sawa T. Information criteria for discriminating among alternative regression models. Econometrica. 1978;46(6):1273-91. [Crossref]
16. Schwarz G. Estimating the dimension of a model. Ann Statist. 1978;6(2):461-4. [Crossref] [Crossref]
17. Hannan EJ, Quinn BG. The determination of the order of an autoregression. J Roy Stat Soc B Met. 1979;41(2):190-5. [Crossref]
18. Hurvich CM, Tsai CL. Regression and time series model selection in small samples. Biometrika. 1989;76(2):297-307. [Crossref]
19. Mallows CL. Some comments on Cp. Technometrics. 1973;15(4):661-75. [Crossref] [Crossref]
20. Hocking RR. A biometrics invited paper. The analysis and selection of variables in linear regression. Biometrics. 1976;32(1):1-49. [Crossref]
21. Golub GH, Heath M, Wahba G. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics. 1979;21(2):215-23. [Crossref]
22. Amemiya T. Selection of regressors. Int Econ Rev. 1980;21(2):331-54. [Crossref]
23. Geweke J, Meese R. Estimating regression models of finite but unknown order. Int Econ Rev. 1981;22(1):55-70. [Crossref]
24. Shibata R. An optimal selection of regression variables. Biometrika. 1981;68(1):45-54. [Crossref]
25. Rissanen J. Stochastic complexity. J Roy Stat Soc B Met. 1987;49(3):223-9. [Crossref]
26. Robinson WS. The statistical measurement of agreement. Am Sociol Rev. 1957;22(1):17-25. [Crossref]
27. Nash JE, Sutcliffe JV. River flow forecasting through conceptual models: part I-A discussion of principles. J Hydrol. 1970;10(3):282-90. [Crossref]
28. Willmott CJ. On the validation of models. Phys Geog. 1981;2(2):184-94. [Crossref]
29. Mielke PW Jr. The application of multivariate permutation methods based on distance functions in the earth sciences. Earth-Sci Rev. 1991;31(1):55-71. [Crossref]
30. Loague K, Green RE. Statistical and graphical methods for evaluating solute transport models: overwiev and application. J Contam Hydrol. 1991;7(1-2):51-73. [Crossref]
31. Watterson LG. Non-dimensional measures of climate model performance. Int J Climatol. 1996;16(4):379-91. [Crossref]
32. Ji L, Gallo K. An agreement coefficient for image comparison. Photogramm Eng Rem S. 2006;72(7):823-33. [Crossref]
33. Fisher RA. A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. Mon Not R Astron Soc. 1920;80(8):758-70. [Crossref]
34. Patry GG, Marino MA. Nonlinear runoff modelling: parameter identification. J Hydraul Eng. 1983;109(6):865-80. [Crossref]
35. Manley RE. Calibration of hydrological model using optimization technique. J Hydraul Div ASCE. 1978;189-202.
36. Stephenson D. Direct optimization of Muskingum routing coefficients. J Hydrol. 1979;41(1-2):161-5. [Crossref]
37. Dust M, Baran N, Errera G, Hutson JL, Mouvet C, Schafer H, et al. Simulation of water and solute transporte in field soils with the LEACHP model. Agr Water Manage. 2000;44(1-3):225-45. [Crossref]
38. Nayak PC, Sudheer KP, Rangan DM, Ramasastri KS. A neuro-fuzzy computing technique for modeling hydrological time series. J Hydrol. 2004;291(1-2):52-66. [Crossref]
39. Willmott CJ, Robeson SM, Matsuura K. Short communication: a refined index of model performance. Int J Climatol. 2012;32(13):2088-94. [Crossref]
40. Joreskog KG. A general method for estimating a linear structural equation system. In: Goldberger AS, Duncan OD, eds. Structural Equation Models in the Social Sciences. 1st ed. New York: Seminar Press; 1973. p.83-112.
41. Tucker LR, Lewis C. A reliability coefficient for maximum likelihood factor analysis. Psychometrika. 1973;38(1):1-10. [Crossref]
42. Wheaton B, Muthén B, Alwin DF, Summers GF. Assessing reliability and stability in panel models. Sociol Methodol. 1977;8:84-136. [Crossref]
43. Bentler PM, Bonett DG. Significance tests and goodness of fit in the analysis of covariance structures. Psychol Bull. 1980;88(3):588-606. [Crossref]
44. Joreskog KG, Sorbom D. Lisrel V. Analysis of Linear Structural Relations by the Method of Maximum Likelihood. 2nd ed. Chicago: International Educational Services; 1981.
45. Hoelter JW. The analysis of covariance structures. Sociol Method Res. 1983;11(3):325-44. [Crossref]
46. Bollen KA. A new incremental fit index for general structural equation models. Sociol Method Res. 1989;17(3):303-16. [Crossref]
47. Bentler PM. Comparative fit indexes in structural models. Psychol Bull. 1990;107(2):238-46. [Crossref] [PubMed]

48. McDonald RP, Marsh HW. Choosing a multivariate model: noncentrality and goodness of fit. Psychol Bull. 1990;107(2):247-55. [Crossref]
49. Browne MW, Cudeck R. Alternative ways of assessing model fit. Sociol Method Res. 1992;21(2):230-58. [Crossref]
50. Steiger JH. Point estimation, hypothesis testing and interval estimation using the RMSEA: some comments and a reply to Hayduk and Glaser. Struct Equ Modeling. 2000;7(2):149-62. [Crossref]
51. Ding J, Tarokh V, Yang Y. Model selection techniques: an overview. IEEE Signal Proc Mag. 2018;35(6):16-34. [Crossref]
52. Pereira HR, Meschiatti MC, Pires RCM, Blain GC. On the performance of three indices of agreement: an easy-to-use r-code for calculating the Willmott indices. Bragantia. 2018;77(2):394-403. [Crossref]
53. Maydeu-Olivares A, Garcia-Forero C. Goodness-of-fit testing. International Encyclopedia of Education; 2010. p.190-6. [Crossref]
54. Dziak JJ, Coffman DL, Lanza ST, Li R. Sensitivity and specificity of information criteria. PeerJ Preprints. 2017:5:e1103v3. [Crossref]
55. Navarro DJ, Myung JJ. Model evaluation. In: Everitt B, Howel D, eds. Encyclopedia of Statistics in Behavioral Science. 1st ed. New York: John Wiley & Sons; 2005. p.1239-42. [Crossref]
56. Rakotoasimbola E, Blili S. Measures of fit impacts: application to the causal model of consumer involvement. Int J Market Res. 2019;61(1):77-92. [Crossref]
57. Willmott CJ, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim Res. 2005;30(1):79-82. [Crossref]
58. Willmott CJ, Matsuura K, Robeson SM. Ambiguities inherent in sums-of-squares-based error statistics. Atmos Environ. 2009;43:749-52. [Crossref]
59. Lin TH, Dayton CM. Model selection information criteria for non-nested latent class models. J Educ Behav Stat. 1997;22(3):249-64. [Crossref] [Crossref]
60. Nadif M, Govaert G. Clustering for binary data and mixture models--choice of the model. Appl Stoch Model D A. 1998;13(3-4):269-78. [Crossref]
61. McDonald RP. An index of goodness-of-fit based on noncentrality. J Classif. 1989;6(1):97-103. [Crossref]
62. Guse B, Pfannerstill M, Gafurov A, Kiesel J, Lehr C, Fohrer N. Identifying the connective strength between model parameters and performance criteria. Hydrol Earth Syst Sci. 2017;21(11):5663-79. [Crossref]