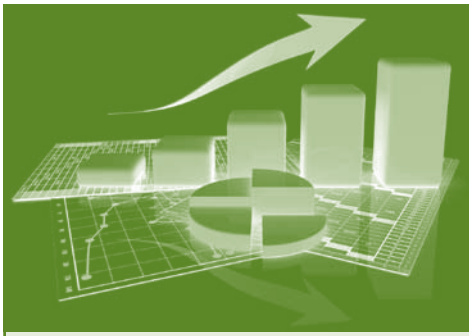
DOI: 10.5455/jmood.20130624120840



Doğrusal Regresyon Analizi

Selim Kılıç[1]

ÖZET:
Doğrusal regresyon analizi

Doğrusal regresyon y olarak isimlendirilen sayısal bir bağımlı değişkenle x olarak ifade edilen bir veya daha fazla bağımsız değişken arasındaki ilişkiyi modelleme yaklaşımıdır. Regresyon modelindeki bağımsız değişken sayısı bir ise model basit doğrusal regresyon olarak tanımlanır. Modeldeki açıklayıcı bağımsız değişken sayısı birden fazla ise çoklu doğrusal regresyon olarak isimlendirilir. Doğrusal regresyonda bağımlı değişken sayısal bir değişken olmak zorundadır. Regresyon modelindeki her bir bağımsız değişkenin 10 katı olgu olması durumunda regresyon model oluşturulması önerilir. İstatistiksel olarak anlamlı bir regresyon analizi bağımlı ve bağımsız değişkenler arasında nedensel ilişki varlığını göstermez.
Anahtar sözcükler: doğrusal, regresyon, nedensellik

ABSTRACT:
Linear regression analysis

Linear regression is an approach to modeling the association between a numeric dependent variable y and one or more independent variables denoted X. The case of one explanatory variable in regression model is called simple linear regression. For more than one explanatory variable, then the model is called multiple linear regression. The dependent variable should be a numeric variable in linear regression. It is recommended at least 10 times as many cases as the number of independent variables in regression model. And a statistically significant regression analysis does not imply causal relationship between independent and dependent variables.
Key words: linear, regression, causation

Journal of Mood Disorders 2013;3(2):90-2

[1]MD, Gülhane Askeri Tıp Fakültesi, Ankara-Türkiye

Yazışma Adresi / Address reprint requests to: Selim Kılıç, Gülhane Askeri Tıp Fakültesi, Ankara-Türkiye

Elektronik posta adresi / E-mail address: drselimkilic@gmail.com

Kabul tarihi / Date of acceptance: 24 Haziran 2013 / June 24, 2013

Bağıntı beyanı:
S.K.: Yazar bu makale ile ilgili olarak herhangi bir çıkar çatışması bildirmemiştir.

Declaration of interest:
S.K.: The author declare that they have no conflict of interests regarding the content of this article.

GİRİŞ

Doğrusal regresyon analizi belirlenmek istenen değişkenden daha kolay veya daha erken saptanabilen değişken(ler)den yola çıkarak belirlenmek istenen değişkeni tahmin eden bir model oluşturmaktır (1). Örneğin Hamilton depresyon puanını, hastaya ilgili ölçek uygulamadan hastanın depresyon şiddeti ile ilişkili olabileceği bilinen diğer değişkenlerle (sözgelimi yaş, kortizol düzeyi, vücut kitle indeksi, kan BDNF düzeyi,...vb) tahmin etmek amaçlanır.

Bilinen normal dağılan sayısal bir değişkenden bilinmeyen, aralarında ilişki olan bir başka normal dağılan sayısal değişkeni tahmin için uygulanırsa basit doğrusal regresyon (simple linear regression), birden fazla değişkenden yararlanarak bir değişkeni tahmin etmek amacıyla modelleme yapıldığında ise "çoklu doğrusal regresyon" (multiple linear regression) olarak tanımlanır (1-4).

Bağımlı değişken (y) ile bağımsız değişken(ler) (x,..) arasındaki ilişkiyi inceler. **Y=a+bX veya Y=B0+B1X** olarak formüle edilir. Değişkenlerden birinin değeri bilindiğinde diğerinin değeri bulunur. Eğer model için bulunan p değeri <0.05 ise regresyon katsayısı 0'dan farklıdır yani iki değişken arasındaki ilişki istatistiksel olarak önemlidir, iki değişken arasında doğrusal bir ilişki vardır. Buna karşılık modelin uygunluk göstergesi R2 ifade edilir ve R2 değeri 1'e ne kadar yakınsa model o kadar iyidir (1-4).

Formülde "a" olarak gösterilen değer, doğrunun y eksenini kestiği nokta (y-intercept, y-kesişim) analitik yöntemde sabit hata ölçüsüdür.

Buna karşılık "b" olarak gösterilen değer regresyon katsayısı (eğim) ise analitik yöntemde oransal (proporsiyonel) hata ölçüsüdür. Bağımsız değişkendeki bir birimlik değişme (artma veya azalma) olduğunda, bağımlı değiş-

kende meydana gelecek ortalama değişim miktarını gösterir (1,3,4).

Doğrusal regresyonda belirlenmek istenen değişken yani bağımlı değişken sürekli veya sıralı sayısal veriler olmalıdır; kategorik değişken olmamalıdır. Buna karşılık bağımsız değişkenlerin de sayısal olması tercih edilir ama bazı durumlarda cinsiyet gibi kategorik de olabilir (1,2,4).

Basit regresyon analizinde her iki değişkenin dağılımı da normal olmalıdır. İlişki katsayısı gibi regresyon katsayısı da pozitif veya negatif olabilir. İlişki katsayısı -1 ile +1 arasında değişirken, regresyon katsayısı her değeri alabilir. Regresyon katsayısı değişkenin ölçüldüğü birimden etkilenir. Sözgelimi hastanın boyunun metre veya santimetre olarak veri tabanında yer alması boya ait katsayısının farklı bulunmasına neden olur. Regresyon analizi aşırı uç değerlerden ve dağılım özelliklerinden ciddi olarak etkilenir. Yine çok önemli bir husus **regresyon modelinin istatistiksel olarak önemli bulunması** tek başına **neden-sonuç ilişkisini açıklamaz (1,4).**

Birden çok bağımsız değişken varlığında ise ÇOKLU REGRESYON yapılabilir. Aşağıda çoklu regresyon modeline bir örnek verilmiştir. Burada β değeri sabit değer ve modelde yer alan değişkenlere ait katsayıları gösterirken, ε ise tesadüfi hata (rezidüeli) göstermektedir. Rezidüel örneklemdeki bireylere ait gözlenen değerle ile bu bireyler için modelden belirlenen değerler arasındaki farktır (1,3).

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_p x_{ip} + \varepsilon_i$$

Bağımsız değişkenlerin; regresyon katsayılarının p değerlerine göre, eşitliğe alınıp alınmamalarına karar verilir. "p" değeri 0.05 ten büyükse eşitliğe alınmaz. Buna karşılık klinik olarak önemli olduğu bilinen, modelde olması gerektiği araştırmacı tarafından gerekli olduğu değerlendirilen değişken(ler) bu değişkenler için bulunan p değerleri >0.05 olsa bile modele alınır ve gerekçesi gereç yöntemde istatistiksel analiz kısmında açıklanır.

Modele alınacak değişkenler arasında dikkat edilmesi gereken önemli bir durum da "çoklu bağıntı" (multicollinearity) kavramıdır. Çoklu bağıntı bir model içerisinde bağımsız değişken olarak birbiriyle yüksek düzeyde ilişkili değişkenler olmasıdır. Bağımsız değişkenler arasında böyle bir ilişkinin olması, değişkenlerden birinin modele ek bir katkı getirmediğine dikkat çeker (1,2,5). Bu durum, bağımsız (yordayıcı) değişkenler arasındaki ilişki katsayılarının mutlak değerinin 0.80'den büyük bulunmasıyla belirlenebilir. Çoklu bağıntı belirlendikten sonra araştırmanın kuramsal temellerini dikkate alınarak bu değişkenlerden sadece biri analize dahil edilip, diğeri/diğerleri analiz dışında tutulmalıdır. Hangi değişken modelde kalmalı sorusunun yanıtı ise öncelikle "klinik olarak önemli olan", bu değişkenle ilgili sorun varsa da "daha kolay ve güvenilir olarak toplanabilen veya ölçülebilen değişken"dir (5).

Bağımsız değişken sayısı fazla olduğunda çeşitli yöntemler yardımıyla modele katkısı en fazla olan daha az sayıdaki değişken veya değişkenler belirlenebilir. Söz konusu yöntemler arasında;

adım adım regresyon yöntemi (stepwise),

ileriye doğru seçim (forward selection),

geriye doğru çıkarma (backward elemination) yöntemleri vardır (1,3,4). SPSS istatistik paket programında doğrusal regresyon analizinin nasıl yapılacağı ve sonuçlarının nasıl yorumlanacağına dair ayrıntılara bazı web adreslerinden ve kaynaklardan ulaşılabilinir (5-10).

Model oluştururken örneklemedeki gözlem sayısının (n), modele dahil edilen bağımsız değişken sayısının en az 5 katı kadar olması, ideal olarak ise, gözlem sayısının bağımsız değişken sayısının 10-20 katı kadar olması belirtilmektedir (1,2,4). Çok önemli bir husus modelin evrene genellenebilmesi için, çalışılacak örneklemin evren için genelleme yapılabilecek yani evreni temsil etmesini sağlayacak uygun yöntemler kullanılarak seçilmesidir (1).

Özet olarak çoklu lineer regresyon; iki veya daha fazla belirleyici (bağımsız) değişkenin bağımlı değişkendeki değişimi açıklamasında kullanılır. Bu değişkenler arasındaki doğrusal ilişkiyi saptayabiliriz fakat altta yatan durumun nedensel bir durum olduğundan kesin olarak emin olamayız.

Kaynaklar:

1. Alpar R. Basit Doğrusal Regresyon Çözümlemesi. Spor, Sağlık ve eğitim Bilimlerinden Örneklerle Uygulamalı İstatistik ve Geçerlik-Güvenirlik. Detay Yayıncılık, Ankara, 2010, 285-304.
2. Kirkwood BR, Sterne JAC. Regression Modelling. Medical Statistics. Blackwell Science. 2003. Australia, 315-342.
3. Pagano M, Gauvreau K. Simple Linear Regression. Principles of Biostatistics . Duxbury Press, 1993, USA, 379-424.
4. Dawson B, Trapp RG. Statistical Methods for Multiple Variables. Basic & Clinical Biostatistics. Lange Medical books/McGraw Hill Medical Publishing Division, 2001, USA, 236-242.
5. Hayran M, Hayran M. Doğrusal Regresyon Analizi, Sağlık Araştırmaları İçin Temel İstatistikler. Omega Araştırma Yayınları. Ankara, 2011, 333-353.
6. https://statistics.laerd.com/spss-tutorials/linear-regression-using-spss-statistics.php. Erişim tarihi 12.06.2013.
7. https://wikis.uit.tufts.edu/confluence/display/SSSI/Simple+Linear+Regression+-+Output. Erişim tarihi 12.06.2013.
8. http://academic.udayton.edu/gregelvers/psy216/SPSS/reg.htm. Erişim tarihi 12.06.2013.
9. http://www.ats.ucla.edu/stat/spss/webbooks/reg/chapter1/spssreg1.htm. Erişim tarihi 12.06.2013.
10. http://www.unt.edu/rss/class/Jon/SPSS_SC/Module9/M9_Regression/SPSS_M9_Regression2.htm. Erişim tarihi 12.06.2013