ÇOKLU DOĞRUSAL REGRESYONDA MODEL SEÇİMİNDE GENELLEŞTİRİLMİŞ TOPLAMSAL MODELLERİN KULLANIMI

USING OF GENERALIZED ADDITIVE MODEL FOR MODEL SELECTION IN MULTIPLE LINEAR REGRESSION

Talat ŞENEL[1*], Mehmet Ali CENGİZ[2], Nurettin SAVAŞ[3], Yüksel TERZİ[2]

[1]Türkiye İstatistik Kurumu, Samsun
[2]Ondokuz Mayıs Üniversitesi, Fen-Edebiyat Fakültesi, İstatistik Bölümü, Samsun
[3]Erzincan Üniversitesi, Fen-Edebiyat Fakültesi, Matematik Bölümü, Erzincan

Geliş Tarihi: 24.08.2009 **Kabul Tarihi:** 04.09.2009

ÖZET

Çoklu doğrusal regresyon yaygın olarak kullanılan istatistiksel yöntemlerden birisidir. Varsayımlar sağlandığında oldukça güçlü bir araçtır. Bu varsayımlardan biri de bağımlı değişkenler ile açıklayıcı değişkenler arasındaki ilişkinin doğrusal, polinomial veya üstel gibi bir bilinen matematiksel fonksiyona sahip olmasıdır. Ancak çoğu uygulamalarda tanımlı böyle bir fonksiyon bulunamayabilir veya bu ilişki kolayca tanımlanamayabilir. Genelleştirilmiş Toplamsal Modeller (GAM), var olan ilişkileri ortaya çıkarmak için tanımlı bir fonksiyonla, parametrik olmayan bir düzleştiriciyi yer değiştirerek bu varsayımı esnetir. GAM çoklu regresyonda model seçiminde de kullanılabilir. Bu çalışma da, çoklu doğrusal regresyon da model seçiminde GAM'ın kullanımı üzerine odaklanılmaktadır.

Anahtar Sözcükler: Çoklu Regresyon; Genelleştirilmiş Toplamsal Modeller, Kübik düzleştirici, Hava kirliliği.

ABSTRACT

Multiple linear regression analysis is one of the most widely used statistical techniques. It is a powerful tool when its assumptions are met, including that the relationships between the predictors and the response are well described with a defined mathematical function (e.g., straight-line, polynomial, or exponential). In many applications, however, the reliance on a defined mathematical function is limiting. Many phenomena do not have a relationship that can be easily defined. Generalized additive models (GAM) enable us to relax this assumption by replacing a defined function with a non-parametric smoother to uncover existing relationships. GAM can be

[*] Sorumlu yazar: tl_senel@hotmail.com

used for model selection in multiple linear regression. This study focuses on GAM for model selection in multiple linear regression.

Key words: Multiple linear regression, Generalized Additive model, Cubic splayn, Air pollution.

1.GİRİŞ

Doğrusal regresyon modellerinde istatistiksel analiz yöntemleri bağımlı değişkenin normal dağıldığı varsayımına dayanmaktadır. Bağımlı değişkenin sürekli olmadığı durumları analiz etmek için de genelleştirilmiş modellerde bulunmaktadır. Bu durumlarda bağımlı değişken sürekli değildir. Ayrıca değişkenleri sürekli olup da normal dağılım göstermeyen verilerde söz konusu olabilir. Bu tür verilerin analizine imkân verecek geliştirilmiş modeller Genelleştirilmiş Doğrusal Modellerdir (GLM).

GLM, ilk kez Nelder ve Wedderburn (1972) tarafından ileri sürülmüştür. Çok geniş uygulama alanlarında kullanılmıştır. Bu alanlardaki ilk detaylı kitaplar McCullagh ve Nelder (1989), Aitkin ve Ark (1989) ve Dobson (1990) tarafından yazılmıştır. Cengiz ve Percy (2001) genelleştirilmiş doğrusal modellerin genel bir özetini vermektedir.

Regresyon modelleri farklı değişkenlerin etkileşimli davranışını anlamak için tahmin, sınıflandırma kuralları ve veri analitik araçlarını sağlayarak pek çok uygulama alanında kullanılmasında önemli bir role sahiptir. Dikkat edilecek olursa basit olmasına rağmen doğrusal modeller, sık sık gerçek yaşam etkilerinde genellikle doğrusal olmadığı için başarısızlığa neden olurlar. Daha esnek istatistiksel modeller doğrusal olmayan regresyon etkilerini tanımlamak için kullanılabilmektedirler. Bu modeller de; Genelleştirilmiş Toplamsal Modeller (GAM) olarak adlandırılırlar.

Toplamsal Modellerin her türlü bağımlı değişken için genelleştirilmiş hali Hastie & Tibrishani (1990) ve Hastie (1991) tarafından önerilmektedir. Bu durum da Genelleştirilmiş Toplamsal Modeller (GAM) olarak adlandırılır. Bu modeller, bilinen GLM'nin Toplamsal Modellere modifiye edilmiş halidir. Bağımlı değişkenin ortalamasının bir toplamsal tahmin ediciye bağlı olduğu bu modellerde doğrusal olmayan bir link fonksiyonu kullanılır. GLM'de

olduğu gibi GAM, bağımlı değişkenin dağılımının üstel dağılımlar ailesinden olmasını ister. GAM'la GLM arasındaki tek fark GAM'ın doğrusal tahmin edici olarak bilinmeyen düzleştirici fonksiyonları kullanmasıdır.

Çoklu regresyon analizin temel varsayımlarından biri olan bağımlı değişken ile açıklayıcı değişkenler arasındaki varsayımın doğrusal gibi bilinen bir matematiksel forma sahip olması varsayımı genelde incelenmez. Bu da bazı açıklayıcı değişkenlerin istatistiksel olarak anlamsızlığına neden olabilir. Doğru bir regresyon modeliyle istatistiksel olarak anlamsız olan açıklayıcı değişkenler anlamlı hale gelebilir. Bu çalışmada, GAM kullanılarak Samsun iline ait hava kirliliğine ilişkin verilere yukarıda bahsedilen çoklu regresyon varsayımları incelenerek daha doğru çoklu regresyon modellerin seçimi yapılmıştır.

2. DOĞRUSAL REGRESYON VE GENELLEŞTİRİLMİŞ TOPLAMSAL MODELLER

Y bir bağımlı tesadüfi değişken ve X_1, X_2, ... , X_p açıklayıcı değişkenler olsun. Bir regresyon işlemi, X_1, X_2, ... , X_p değerleri verildiğinde Y'nin beklenen değerini tahmin etmede kullanılan bir istatistik yöntem olarak ifade edilebilir. Standart çoklu doğrusal regresyon, koşullu beklenti ile ifade edilen

$$E(Y|X_1, X_2, \cdots, X_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \qquad (1)$$

bir doğrusal form varsayar. Verilen bir örnek için, $\beta_0, \beta_1, \cdots, \beta_p$ tahminleri genellikle en küçük kareler yöntemiyle elde edilir.

Toplamsal model,

$$E(Y|X_1, X_2, \cdots, X_p) = s_0 + s_1(X_1) + s_2(X_2) + \cdots + s_p(X_p) \qquad (2)$$

koşullu beklentisinin modellenmesiyle doğrusal modelleri genelleştirir. Burada $s_i(X)$, i = (1,2, ... , p) düzleştirme fonksiyonlarıdır. Tahmin edilebilir olması için düzleştirme fonksiyonları s_i 'ler $E[s_j(X_j)] = 0$ gibi standart koşulları sağlaması gerekir. Bu fonksiyonlar parametrik bir forma sahip değillerdir ancak tahminleri parametrik olmayan bir biçimde elde edilir.

Klasik doğrusal modeller ve toplamsal modeller pek çok istatistiksel veri analizinde kullanılabilirken, bu modellerin uygun olmadığı durumlarda söz konusudur. Örneğin, normal dağılım bağımlı değişkenlerin kategorik olduğu durumlarda uygun değildir veya bağımlı değişken ile açıklayıcı değişken arasındaki ilişki bilinen parametrik bir formda olmayabilir. Genelleştirilmiş Toplamsal Modeller bu tür zorlukların aşılmasını sağlayabilir. Hem bağımlı değişkenin normal dağılım göstermediği hem de bağımlı değişkenle açıklayıcı değişken arasındaki ilişkinin parametrik (doğrusal yada kübik olmadığı durumlar) olmadığı durumlarda, GAM modellemeye esneklik getirir.

Genelleştirilmiş Doğrusal modellere benzer şekilde GAM'da bir tesadüfi bileşen, bir toplamsal bileşen ve bir de bu iki bileşeni birleştiren link fonksiyonu içerir. Bağımlı değişken Y, tesadüfi bileşen aşağıdaki gibi forma sahip üstel dağılımlar ailesinden bir yoğunluğa sahip olduğu varsayılır.

$$f_Y(y;\theta;\phi) = \exp\left\{\frac{y(\theta) - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \qquad (3)$$

Burada θ doğal parametre ve ϕ ölçek parametresi olarak isimlendirilir. Bağımlı değişken ortalaması, bir g link fonksiyonu kullanarak X_1, X_2, ... , X_p açıklayıcı değişkenleriyle ilişkilendirilir. Toplamsal bileşen (4) deki gibi ifade edilir.

$$\eta = s_0 + \sum_{i=1}^{p} s_i(X_i) \qquad (4)$$

Burada $s_1(\cdot)$, ... , $s_p(\cdot)$ düzleştirme fonksiyonlarıdır ve μ ve η arasındaki ilişki $g(\mu) = \eta$ ile tanımlıdır.

Bir düzleştirme, bir veya birden fazla açıklayıcı değişkenin ölçümlerinin X_1, ... , X_p 'nin bir fonksiyonu olarak bir bağımlı değişken Y'nin trendini özetlemede kullanılan bir araçtır. Y'nin kendisinden daha az değişkene sahip trendin bir tahminini verir. Bir düzleştiricinin en önemli özelliği parametrik olmayan yapısıdır. Y 'nin X_1, ... , X_p üzerine bağımlılığı için kesin bir form varsaymaz. Farklı düzleştiriciler söz konusu olmasına rağmen bu çalışma kübik düzleştirici zinciri ile sınırlandırılmıştır. Kübik düzleştirici zinciri

olarak isimlendirilen düzleştirici basit bir optimizasyon probleminin çözümüdür. Bu problem aşağıdaki gibi ifade edilebilir. İkinci dereceden sürekli türevlere sahip bütün $\eta(x)$ fonksiyonları arasından, (5) formülünde verilen cezalı en küçük kareler ifadesini minimum yapanı bulmaktır.

$$\sum_{i=1}^{n}(y_i - \eta(x_i))^2 + \lambda\int_a^b(\eta''(t))^2\,dt \qquad (5)$$

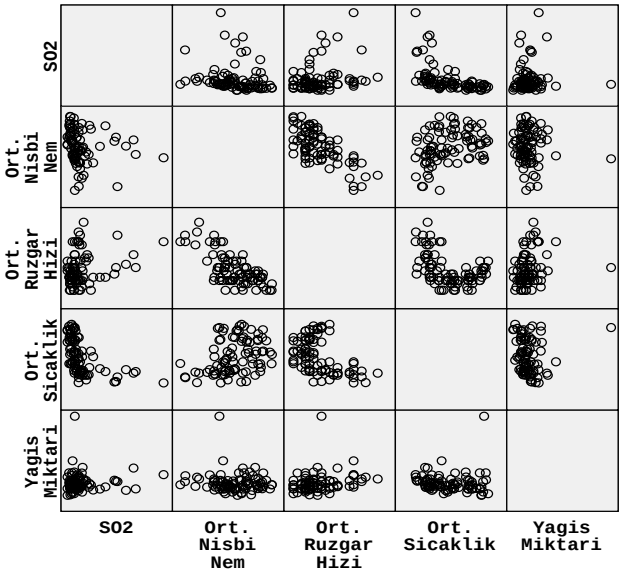
Burada λ bir sabit ve $a \le x_1 \le \cdots \le x_n \le b$ dir. İkinci terim fonksiyondaki eğriliği cezalandırırken, birinci terim veriye olan yakınlığı ölçer. Açık ve tek bir minimum değer olduğu gösterilebilir ve minimum yapan x_i 'nin tek değerlerinde düğüm noktalarına sahip bir kübik zincirdir. λ düzleştirme parametresidir. λ'nın büyük değerleri daha düzgün eğriler verir.

3. UYGULAMA

Kükürt dioksit (SO_2) kirliliği; kükürt içeren fosil yakıtların yanması ile şehirsel ısınmada ve bazı endüstriyel süreçlerin sonucunda bacalardan atılan kirliliklerden oluşmaktadır (Bayram,2005). SO_2 ölçümlerindeki değişim Nispi Nem, Ortalama Rüzgâr Hızı, Ortalama Sıcaklık ve düşen yağış miktarları gibi meteorolojik ölçümler ile ilişkilidir. Bu ilişkiyi ortaya koyan istatistiksel yöntemlerin başında çoklu regresyon gelmektedir. Bu çalışmada ilk olarak çoklu doğrusal regresyon modelleri ve Genelleştirilmiş Toplamsal Modeller verildikten sonra 1998-2008 yılları arasındaki Samsun il merkezindeki hava kirliliği ve meteorolojik verilere çoklu doğrusal regresyon ve Genelleştirilmiş Toplamsal Modeller karşılaştırmalı olarak uygulanmıştır.

Uygulama verisi Devlet Meteoroloji İşleri istasyonundan ve TUİK veri tabanından elde edilmiştir. Temel değişkenler 1998 ile 2008 yılları arasındaki aylık ortalama olarak alınan Nispi Nem (X_1) , Ortalama Rüzgâr Hızı (X_2), Ortalama Sıcaklık (X_3), Aylık Yağış Miktarları (X_4) ve Samsun il merkezi için hava kirliliğine neden olan SO_2 (Y) değerleri gibi sürekli değişkenler alınmıştır. Verimize ilk olarak bağımlı değişken ve açıklayıcı değişkenlerin arasındaki yapıyı yüzeysel olarak anlayabilmek için çoklu serpilme diyagramı

yapılmıştır.



Şekil 1. Serpilme Diyagramı

İlk olarak açıklayıcı değişkenler olan Nispi Nem (X_1) , Ortalama Rüzgar Hızı (X_2), Ortalama Sıcaklık (X_3), Aylık Yağış Miktarları (X_4) değişkenlerinin hava kirleticisi olan SO_2 üzerine etkilerini incelemek amacı ile çoklu regresyon modeli uygulanmıştır. Bu amaçla SAS programı içerisinde PROC GENMOD kullanılmıştır. GENMOD işlemi değişkenlere Genelleştirilmiş Doğrusal Modeller uygular. Parametre vektörünün tahmini için en çok olabilirlik tahmin yöntemini kullanır. Genelde bu tahminler için açık çözüm formları yoktur. Dolayısıyla nümerik çözüm için yöntemler kullanılır. Bu analizin amacı SO_2 ölçümleri ile açıklayıcı değişkenler arasındaki ilişkiyi ortaya koymaktır.

Bu analiz ile SO_2 oranı ile Ort. Nispi Nem, Ort. Rüzgar Hızı, Ort. Sıcaklık ve Yağış Miktarı arsındaki çoklu analizi yaparak parametre tahminlerini gerçekleştirmektir. Parametre tahminleri Tablo 1 'deki gibidir.

Tablo 1. Parametre Tahminleri Tablosu

Parametre tahminleri analizi

Parametre	SD	Tahmin	S. Hata	Wald 95% Güven aralığı		Ki-kare	p
Sabit	1	216.710	135.690	-49.237	482.657	Şub.55	0.1102
Nispi NEM (X_1)	1	0.0530	0.1062	-0.1551	0.2611	0.25	0.6179
Ort.Rüzgar Hızı (X_2)	1	70.593	30.917	0.9996	131.190	May.21	0.0224
Ort.Sıcaklık (X_3)	1	-0.9147	0.2268	-13.592	-0.4701	16.26	<.0001
Aylık Yağış Miktarı (X_4)	1	0.0468	0.0555	-0.0620	0.1557	0.71	0.3991

Tablo 1'de modelimizin denklemi;

$$SO_2 = 21.671 + 0.0530X_1 + 7.0593X_2 - 0.9147X_3 + 0.0468X_4$$

olarak tahmin edilir. Burada p=0.05 anlamlılık seviyesinde SO_2 yi açıklamada Ort. Nispi Nem ve Aylık Yağış Miktarı'nın anlamsız faktörler olduğunu, Ort. Rüzgar hızı ve Ort. Sıcaklığın ise anlamlı faktörler olduğu açıktır.

Bu süreçte SO_2 ile açıklayıcı değişkenler arasında doğrusal bir ilişki olduğu varsayılmaktaydı. Acaba bu varsayım ne kadar doğrudur? Bunu ölçmek için aynı veriye kübik zincir düzleştiricili Genelleştirilmiş Toplamsal Modeller uygulanabilir. Bu amaca yönelik olarak SAS PROC GAM kullanılmıştır. PROC GAM süreci parametrik olmayan bir yöntemle açıklayıcı değişkenlerin bir bağımlı değişkeni nasıl açıkladığını ortaya koyan bir yöntemdir.

Model olarak bir toplamsal model kullanılmış ve her bir açıklayıcı değişken için bir tek değişkenli düzleştirici zincir uygulanmıştır. Dolayısıyla her bir açıklayıcı değişken 3 serbestlik derecesine sahip bir kübik zincir kullanılarak fit edilmiştir. Bu 3 serbestlik derecesinin bir tanesi uyumun doğrusal kısmı için alınırken kalan ikisi uyumun doğrusal olmayan zincir kısmı için alınmıştır.

İlk olarak, SO_2 ile açıklayıcı değişkenler arasındaki ilişkinin doğrusal olup olmadığı test edilip sonuçlar Tablo 2' de sunulmuştur.

Tablo 2. GAM ile doğrusallık varsayımının incelenmesi

Parametre tahminleri analizi				
Parametre	**Tahmin**	**S. Hata**	**t**	**p**
Sabit	-127.909	1.234.863	-0.10	0.9178
Lineer (X1)	0.17841	0.09664	Oca.85	0.0684
Lineer (X2)	1.413.707	281.368	05.Şub	<.0001 *
Lineer (X3)	-106.094	0.20641	-5.14	<.0001 *
Lineer(X4)	0.05044	0.05054	1.00	0.3211

Tablo 2' den 0.05 anlamlılık seviyesinde Ort.Rüzgar hızı ve Ort.Sıcaklık doğrusal çıkarken, Ort.Nisbi Nem ve Aylık Yağış Miktarları doğrusal olmadığı gözükmektedir. Acaba doğrusal olmayan ilişkiler kuadratik (quadratic) olabilir mi? sorusununa, yine SAS PROC GAM program ile yanıt verilebilir. İlişkinin kuadratik olup olmadığı Tablo 3 de görülebilir.

Tablo 3 GAM ile ilişkinin **kuadratik** olup olmadığının incelenmesi

Düzleştirme Model analizi sapma analizi				
Parametre	**SD**	**Kareler toplamı**	**Ki-kare**	**p**
Zincir(X_1)	300.000	1.805.492.885	149.006	0.0019*
Zincir(X_2)	300.000	263.928.450	21.782	0.5363
Zincir(X_3)	300.000	228.681.523	18.873	0.5961
Zincir(X_4)	300.000	1.689.923.021	139.468	0.0030*

Tablo 3'e göre 0.05 anlamlılık seviyesinde SO_2 ile Ort. Nispi Nem(X_1) ve Aylık Yağış Miktarı (X_4) değişkenleri arasındaki ilişki doğrusal olmayıp kuadratiktir.

Çoklu doğrusal regresyon modeli ve kübik düzleştirici zincirli GAM kullanımıyla, X_2 ve X_3 değişkenlerinin sadece doğrusal formda X_3 ve X_4 değişkenlerinin hem doğrusal hem de kuadratik formda

modele katılması gerektiği ortaya çıkmıştır. Son olarak SO_2 bağımlı değişken, Ort. Nispi Nem(X1) ve Aylık Yağış Miktarı (X_4) açıklayıcı değişkenleri karesel formda, Ort. Rüzgâr hızı ve Ort. Sıcaklık açıklayıcı değişkenleri sadece doğrusal olarak modellenmesi incelenebilir. Bunun için, SAS PROC GENMOD tekrar kullanılırsa Tablo 4'deki sonuçlar elde edilir.

Tablo 4 GAM için parameter tahminleri tablosu

Parametre tahminleri analizi

				Wald 95%			
Parametre	**SD**	**Tahmin**	**S. Hata**	**Güven aralığı**		**Ki-kare**	**p**
Sabit	1	-718.019	365.425	-143.424	-0.1799	Mar.86	0.0494
X_2	1	119.348	33.936	52.836	185.861	Ara.37	0.0004
X_3	1	-10.676	0.2043	-14.680	-0.6672	27.31	<.0001
X_1	1	19.846	0.6016	0.8054	31.637	Eki.88	0.0010
X_1*X_1	1	-0.0079	0.0024	-0.0126	-0.0031	Eki.38	0.0013
X_4	1	-0.4859	0.1528	-0.7854	-0.1864	10.Kas	0.0015
X_4*X_4	1	0.0044	0.0012	0.0021	0.0068	13.81	0.0002

Tablo 4'e göre yeni modelimizin denklemi;

$$SO_2 = -71,8019 + 11,9348X_2 - 1,0676X_3 + 1,9846X_1 - 0,0079X_1^2 - 0,4859X_4 + 0,0044X_4^2$$

olarak bulunur. Görüldüğü gibi tüm açıklayıcı değişkenler 0.05 anlamlılık seviyesinde istatistiksel olarak anlamlıdır.

4. SONUÇ VE TARTIŞMA

Çoklu regresyon analizin temel varsayımlarından biri olan bağımlı değişken ile açıklayıcı değişkenler arasındaki varsayımın doğrusal ya da üstel gibi bilinen bir matematiksel forma sahip olmasıdır. Bu varsayım genelde göz ardı edilir. Buda aslında istatistiksel olarak anlamlı bazı değişkenlerin anlamsız gözükmesine neden olabilir. Açıklayıcı değişkenlerle bağımlı değişken arasındaki ilişkini doğrusal varsayıldığı çoklu doğrusal regresyon modeli

uygulandığında Ort.Nisbi Nem (X_1), Ort.Rüzgar Hızı (X_2), Ort.Sıcaklık (X_3), Ort. Yağış Miktarı (X_4) açıklayıcı değişkenlerinden 0.05 anlamlılık seviyesinde SO_2'yi açıklamada Ort. Nispi Nem ve Aylık Yağış Miktarı'nın anlamsız faktörler olduğu, Ort. Rüzgar hızı ve Ort.Sıcaklığın ise anlamlı faktörler olduğu gözükmektedir. Oysa GAM kullanılarak açıklayıcı değişkenlerle bağımlı değişkenler arasındaki ilişkiler incelendiğinde 0.05 anlamlılık seviyesinde Ort. Rüzgar hızı ve Ort.Sıcaklık doğrusal çıkarken, Ort.Nispi Nem ve Aylık Yağış Miktarları doğrusal olmadığı kuadratik bir ilişkinin söz konusu olduğu anlaşılmaktadır. Bu ilişki yapıları göz önüne uygun çoklu regresyon modeli uygulandığında anlamsız açıklayıcı değişkenler olan Ort. Nispi Nem ve Aylık Yağış Miktarı'nın da oldukça önemli düzeyde anlamlı hale geldiği açıktır.

Sonuç olarak, GAM çoklu doğrusal regresyon modellerin açıklayıcı değişkenler yapısının belirlenmesinde kullanılabilir. Bu çalışmada aradaki ilişkinin kuadratik olup olmadığını görmek amacıyla sadece tek bir düzleştirici zincir ile çalışılmıştır. Alternatif farklı düzleştirici zincirlerin çalışılması farklı yapıların da incelenmesine neden olacaktır. Bu da uygulama çoğu zaman göz ardı edilen bazı varsayımların farkında olunmadan çiğnemesinde ortaya çıkacak modelleme hatalarının minimum indirilmesine neden olacaktır.

KAYNAKLAR

Aitkin, M., Anderson, D., Francis, B., Hinde, J. (1989). *Statistical modelling in GLIM*, Clarendon Press, Oxford.

Bayram, H. (2005). Türkiye'de Hava Kirliliği Sorunu: Nedenleri, Alınan Önlemler ve Mevcut Durum. *Toraks Dergisi*, 2005;6(2):159– 165.

Cengiz, M.A., Percy, D.F. (2001). Mixed Multivariate Generalized Linear Models for Assessing Lower-Limb Arterial Stenoses, *Statistics in Medicine,* 20, 1663-1679.

Dobson, A.J., Barnett, A.G. (2008). *Introduction to Generalized Linear Models,* Third Edition. London: Chapman and Hall/CRC.

Hastie T.J, Tibshirani RJ (1990). *Generalized Additive Models,* New York: Chapman and Hall.

Hastie, T. J. (1991). *Generalized Additive Models, in Statistical Models in S*, ed. J. M. Chambers and T. J. Hastie, Pacific Grove: Wadsworth & Brooks/Cole Advanced Books & Software.

McCullagh, P., Nelder, J.A. (1989). *Generalized Linear Models,* 2nd ed., New York: Chapman and Hall.

Nelder, J., Wedderburn, R. (1972). Generalized Linear Models, *Journal of the Royal Statistical Society,* 370-384.

****