Lojistik Regresyonda Bayesci Model Ortalaması Yaklaşımı

Mehmet Ali CENGİZ[1], Naci MURAT[1], Yüksel TERZİ[2*], Nurettin SAVAŞ[3]

[1]Ondokuz Mayıs Üniversitesi, Fen-Ed. Fak. İstatistik Bölümü, 55139 Samsun
[2]Afyon Kocatepe Üniversitesi, Fen-Ed.Fak. İstatistik Bölümü, 03200 Afyonkarahisar
[3]Erzincan Üniversitesi, Fen-Ed.Fak. Matematik Böl., Erzincan

Özet: Standart istatistiksel metotlar model belirsizliğini ihmal eder. Veri analizcileri olası model sınıfından bir model seçer ve sanki seçilen model veriyi üretmiş gibi işleme devam eder. Bu yaklaşım model seçiminde belirsizliği ihmal ederek istatistiksel çıkarımlar için güven aralıklarını daha geniş tutar ve daha riskli kararlara neden olur. Oysa *Bayesci model ortalaması (BMA)* bu model belirsizliğini göz önüne alan bir yapı sunar. Bu çalışmada *BMA* yaklaşımını sunularak gerçek hayattan bir probleme uygulaması verilmiştir. Uygulamada, BMA yaklaşımının örnek kestirim performansını geliştirdiği görülmüştür.

Anahtar kelimeler: Bayesci yaklaşım, Bayesci model ortalaması, model belirsizliği

Bayesian Model Averaging Approch In Logistic Regression

Abstract: Standard statistical practice ignores model uncertainty. Data analysts typically select a model from some class of models and then proceed as if the selected model had generated the data. This approach ignores the uncertainty in model selection, leading to over-confident inferences and decisions that are more risky than one thinks they are *Bayesian model averaging (BMA)* provides a coherent mechanism for accounting for this model uncertainty.. In this study, we discuss *BMA* approach and present a real life application. In this application, *BMA* provides improved out-of sample predictive performance.

Key words: Bayesian approach, Bayesian model averaging, model uncertainty

* E-mail: yukselt@aku.edu.tr

1. Giriş

Bayesci model ortalaması (BMA) model seçiminde belirsizliği ortadan kaldırmaya yarayan bir yaklaşımdır. Pek çok farklı model ortalamasıyla kullanılabilen BMA yaklaşımı, model belirsizliğini parametre ve kestirimlere ilişkin sonuçlarla birleştirir. Günümüzde, *BMA doğrusal regresyon, genelleştirilmiş doğrusal modeller, Cox regresyon* modelleri, kesikli grafik modelleri gibi pek çok istatistiksel modellemelerde başarıyla uygulanmaktadır. Pek çok çalışmada kestirim performansını artırdığı gözlenmiştir.

BMA temelde modellerin birleştirilmesi fikrine dayanmaktadır. Bu fikrin çok eskilere dayanmasına rağmen kullanımı yenidir. Bernard (1963) ve Bates ve Granger (1969) ön tahminleri birleştirme için metotlar geliştirmişlerdir. Ekonomik ve hava tahminlerine ilişkin pek çok çalışma yapılmıştır. Clemen (1989) de bu çalışmaları detaylı bir şekilde anlatmıştır.

Model ortalamasına ilişkin ilk çalışma Roberts (1965) dir. Roberts (1965) iki uzman (veya model) görüşlerini birleştiren bir dağılım önermiştir. Bu dağılım, iki modelin sonsal dağılımlarının bir ağırlıklı ortalamasıdır ve *BMA*'ya benzerdir. Leamer (1978), bu fikri geliştirerek *BMA* için temel oluşturmuştur. Leamer (1978) , klasik model seçimindeki belirsizliği ortaya atarak, yöntemin bu belirsizliği göz önüne aldığını vurgulamıştır.

George (1999), karar teorisinde *BMA*'nın kullanımını incelerken, Drapper (1995), Chatfield (1995) ve Kass ve Raftery (1995) de *BMA* kullanımını genişletmiştir. Hoeting ve ark. (1999) da *BMA* için özet bir çalışma sunmuştur.*BMA*'nın tahmin performansını ölçmek için pek çok simülasyon çalışması da yapmıştır. Clyde (1999) da loglineer modeller ve Clyde ve George (2000) de wavelet'lar üzerine çalışmış ve *BMA*'nun klasik model seçimi yöntemlerinden daha iyi performans gösterdiğini ifade etmiştir.

BMA'nın performansının gösterimi amacıyla pek çok çalışmada gerçek verilerde uygulanmıştır. Madigan ve Raftery (1994) de grafiksel modeller için, Raftery ve ark. (1995) sağkalım analizi için, Fernandez ve ark. (2001) ve Hoeting ve ark. (2002) doğrusal modeller için, Fernadez ve ark. (2002) binary regresyon için ve Lamon ve Clyde (2000) yarı *parametrik regresyon* modelleri için *BMA* yı farklı gerçek verilere uygulamış ve daha iyi performans gösterdiğini ifade etmişlerdir.

Bu çalışmada *Bayesci Model Ortalaması* yaklaşımı ve gerçekleştirilmesi verilerek, Kalp damar hastalığının risk faktörlerinin belirlenmesi üzerine uygulanmıştır. *Lojistik regresyon modeli* kullanılarak, model seçiminde *BMA* yaklaşımı ile Klasik model seçimi yöntemleri karşılaştırılmıştır.

2. *Bayesci Model Ortalaması*

Bayesci model ortalaması bütün olasılık modellerinin bir alt kümesini seçer. $K=2^p$ kadar alt küme seçebilir ve açıklayıcı değişkenlerin etkileşimini ihmal eder. Bütün çıkarımları ve kestirimleri elde etmek için modellerin sonsal olasılıklarını kullanır.

$$M = \{M_1, M_2, ..., M_K\} \qquad (1)$$

ilgilenilen bütün olası modellerin kümesini göstersin. Δ her bir modelde aynı yoruma sahip ilgilenilen regresyon parametreleri veya tahmin edilecek gelecek değerler gibi bir nicelik olsun. Böylece D verisi bilindiğinde Δ ' nın sonsal dağılımı;

$$P(\Delta \setminus D) = \sum_{k=1}^{K} P(\Delta \setminus M_k, D) P(M_k \setminus D) \qquad (2)$$

Bu her bir M_k modeli altındaki sonsal dağılımların bir ortalamasıdır. Bir M_k modeli verildiğinde Δ ' nın predictive dağılımı;

$$P(\Delta \setminus M_k, D) = \int P(\Delta \setminus \beta^k, M_k, D)\, P(\beta^k \setminus M_k, D)\, d\beta^k \qquad (3)$$

$$\beta^k = (\beta_0, \beta_1, ..., \beta_p)'$$

(3)'daki eşitlik M_k modeli için regresyon parametrelerinin vektörüdür. Burada:
M_k modelinin sonsal olasılığı,

$$P(M_k \setminus D)=\frac{P(D/M_k)P(M_k)}{\sum_{j=1}^{K} P(D/M_j)P(M_j)} \quad (4)$$

M_k modelinin integrallenmiş olabilirliği;

$$P(D/M_k) = \int P(D \setminus \beta^k, M_k) P(\beta^k \setminus M_k) d\beta^k \quad (5)$$

β^k 'nın önsel yoğunluğu

$$P(\beta^k / M_k), \quad (6)$$

$P(D/\beta^k, M_k)$ olabilirlik, $P(M_k)$ M_k'nın optimal olduğuna ilişkin önsel olasılıktır.

2.1. BMA'nın Gerçekleştirilmesi
Uygulamaya ilişkin bir takım zorluklar söz konusudur. Bunlar;

- Modellerin $P(M_k)$ önsel olasılıklarının tespiti
- β^k parametrelerinin önsel dağılımının tespiti
- İntegrallerin hesaplanması (genelde analitik çözüm yoktur)
- Çok sayıda olası model söz konusu iken Δ'nın sonsal dağılımının hesaplanmasıdır.

Modellerin olabilirliğine ilişkin çok az bilgi varsa her birinin gerçekleşme olasılığını eşit almak makul bir seçenektir.

β^k parametreleri için önsel olarak çok değişkenli normal önsel dağılım alınabilir.
Ortalama : ençok olabilirlik tahmincisine ve
Varyans : gözlem için bilgi matrisinin beklenenine eşittir Raftery (1995,1996,1999).

$$P(D/M_k) = \int P(D \setminus \beta^k, M_k) P(\beta^k \setminus M_k) d\beta^k \quad (7)$$

integrali Laplace Metodu kullanılarak yaklaşık olarak hesaplanabilir.

$$\log P(D \setminus M_k)=\log P(D/\hat{\beta}^k, M_k) - P_k \log n + O(1) \quad (8)$$

Burada;
$\hat{\beta}^k$: M_k modeli altında β_k parametre vektörünün sonsal ortalaması
P_k: Model M_k'daki parametre sayısı
n: verideki gözlem sayısı
Bu eşitlik *Schwarz Bayesci* bilgi kriteri olarak da bilinir. P ortak değişkenli bir analiz için model sayısı K oldukça büyük olabilir. Bu problemi aşmak için Madigan ve Raftery (1994) tarafından önerilen The Occam's Window yaklaşımını kullanırız. Bu yaklaşım sadece en yüksek sonsal model olasılıklarına sahip olasılıkları göz önüne alır. Veri verildiğinde bir modelin sonsal olasılığı diğer olası modellerden küçükse, bu model göz ardı edilir. Sadece

$$A = \left[M_k : \frac{maks_l \; P(M_l / D)}{P(M_k / D)} \leq C \right] \qquad (9)$$

kümesine ait modeller hesaplamaya katılır. C'nin 20 alınması genel bir kabuldür (Raftery, 1995, 1996).

3. Uygulama

Kalp damar hastalığı (CAD) ülkemizde ölümcül oranı en yüksek hastalıkların başında gelmektedir. Yüksek kolesterol, sigara içme, hiper tansiyon, insülin seviyesinin yüksekliği, yaş, cinsiyetin erkek olması, şişmanlık, alkol kullanımı, diyabetik bir rahatsızlığın olması ve yakınlarında bu hastalığın olması en temel risk faktörleridir.

Kullanılan veri Erzurum Atatürk Üniversitesi Tıp Fakültesinde toplanılmış olup Balcı ve ark. (2000) tarafından kullanılmıştır. Normal şeker seviyesine sahip erkek hastalar için plazma insülin seviyeleri ile CAD'ın angiographical yoğunluğu arasındaki ilişki araştırılmıştır. Bu çalışmada 32'si kontrol diğerleri hasta grubu olan 101 bireyden oluşan veriye BMA metodu uygulanarak CAD için risk faktörleri lojistik regresyonla ortaya konulmuştur. Elde edilen ölçümler Tablo 1'de özetlenmiştir;

Tablo1. Değişkenler ve tanımlamaları

DEĞİŞKEN	TANIM
Cad	Kalp Damar Hastalığı (1,0)
Yaş	Bireylerin Yaşları
İnsülin	Açlık İnsülin Seviyesi
Kolesterol	Ortalama Kolesterol Seviyesi
Soy geçmiş	Yakınlarında CAD olup olmadığı
Sigara	Sigara İçip İçmediği
Tansiyon	Alkol Kullanıp Kullanmadığı
Hipertansiyon	Hipertansiyon Olup Olmadığı

Etkileşim ihmal edilerek sadece ana etkiler kullanılmıştır. 8 olası açıklayıcı değişkenin bütün olası kombinasyonları gözden geçirildi. Başlangıçta 256 olası model söz konusu idi. Bic.logit programı kullanılarak Occam's window metoduyla 256 olası model 4'e indirgenmiştir. Seçilen modeller ve sonsal model olasılıkları Tablo 2'de verilmiştir.

Tablo 2. Farklı Model Seçme Kriterlerine Göre Oluşturulan Modeller

DEĞİŞKEN	BMA				ADIMSAL ENTRY
	1	2	3	4	5
Yaş	√	√	√	√	
Açlık İnsülin	√	√	√	√	√
Kolestrol	√	√	√	√	
Soygeçmiş	√				√
Sigara	√	√	√		√
Alkol	√				√
Hipertansiyon	√				√
Sapma	56,327	59,001	59,012	59,315	58,998
Sonsal Model Olasılığı	0,812	0,473	0,451	0,376	

En yüksek sonsal olasılığa sahip model (% 81,2) 7 açıklayıcı değişken içermektedir. Oysa adımsal (entry) yöntem 6 değişken içermektedir. Kolestrol seviyesini modele katmamaktadır. Tablo 3'de BMA'nın seçtiği en iyi yöntemle klasik adımsal (entry) yöntemin sonuçlarını karşılaştırmaktadır.

Tablo 3. BMA ve Adımsal (Entry) Metodu sonuçlarının karşılaştırılması

	BMA		ADIMSAL (ENTRY)	
DEĞİŞKEN	**Ortalama**	**Yüzde**	**Katsayı**	**P değeri**
Yaş	0,081	76,4	0,085	0,068
Açlık İnsülin (log)	0,563	100	0,579	0,06
Kolestrol	0,028	87,4	0,023	0,659
Soygeçmiş	1,631	75,5	1,584	0,023
Sigara	2,135	100	2,243	0,006
Alkol	1,953	68,5	2,053	0,09
Hipertansiyon	1,771	92,6	1,871	0,015

4. Sonuç

Teoride *BMA* yöntemi model seçiminde iyi bir performans göstermektedir. Teorideki bu durum uygulamalı bir çalışmayla da gösterilmiştir. Bu çalışmada *BMA*'nın *lojistik regresyonda* model seçimindeki performansı üzerinde durulmuştur. Araştırılması gereken açık sorular vardır. Bunlar: farklı önsellerin seçimi, kestirim performansının geliştirilmesi ve farklı modellemelere uygulanmasıdır. Basit yada çoklu regresyon modellerinde, çok değişkenli regresyon modellerinde, farklı genelleştirilmiş lineer modellerde (*Loglineer modeller* gibi), *sağkalım analizinde* ve grafiksel modellerde de uygulanabilir. Markov Zinciri, Monte Carlo (MCMC) yöntemleri de *Bayesci hipotez testlerinde* ve *Bayes faktörü* hesaplamaları için incelenebilir. *BMA* da çok büyük sayıda modellerle çalışıldığı için her bir model için bir önsel bilgi seçimi zorunluluğu vardır. Bu zorlukları aşmada literatürde yapılan çalışmaları olmasına rağmen halen yeni çalışmalara ihtiyaç vardır. Benchmark önsel seçimi yapılabilir.

Kaynaklar

1. Bates, J. M. and Granger, C. W. J., **The Combination of Forecast**, Operational Research Quarterly, 20, 451-468, (1969).
2. Chatfield, C., **Model Uncertainty, Data Mining, and Statistical inference (With Discussion)**, J. Roy. Statist. Soc. Ser. A, 158, 419-466, (1995).
3. Clemen, R. T., **Combining Forecasts: A Review and Annotated Bibliography**, Internat. J. Forecasting, 5, 559-583, (1989).
4. Clyde, M. A., **Bayesian Model Averaging and Model Search Strategies (With Discussion) İn Bayesian Statistics,** 6, Eds. J. M. Bernardo Et Al., Oxford University Pres. Pp. 157-185, (1999).
5. Clyde, M. A., and George, E.I., **Flexible Empirical Bayes Estimation for Wavelets,** Journal Of The Royal Statistical Society. Ser. B, 62, 681-698, (2000).
6. Draper, D., **Assessment And Propagation of Model Uncertainty**, J. Roy. Statist. Soc. Ser. B, 57, 45-97, (1995).
7. Fernadez, C., Ley, E., and Steel, M.F.J., **Benchmark Priors for Bayesian Model Averaging**, Journal Of Econometrics, 1000, 381-427, (2001a).
8. Fernadez, C., Ley, E., and Steel, M. F. J., **Bayesian Modelling of Catch in a North-West Atlantic Fishery**, Applied Statistics, 51, 257-280, (2002).
9. George, E. I., **Bayesian Model Selection**, In Encyclopaedia of Statistical Science Update 3. Wiley, New York, to Appear, (1999).
10. Hoetıng, J. A., Madıgan, D. M., Raftery, A. E. and Volinsky, C. T., **Bayesian Model Averaging: a Tutorial (With Discussion)**, Stat. Sci., 14, 382-401, (1999).

11. Hoetıng, J. A., Raftery, A. E., and Madıgan D., **Bayesian Variable and Transformation Selection in Linear Regression**, Journal Of Computational and Graphical Statistics, 11, 485-507, (2002).

12. Kass, R. E. and Raftery, A. E., **Bayes Factors**, J. Amer. Statist. Assoc., 90, 773-795, (1995).

13. Leamer, E. E., **Specification Searches**, Wiley, New York, (1978).

14. Madıgan, D. and Raftery, A. E., **Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window**, Journal Of The American Statistical Association, 89, 1335-1346, (1994).

15. Raftery, A. E., Madıgan, D. and Hoetıng, J., **Bayesian Model Avaraging for Linear Regression Models,** J. Amer. Statist. Assoc., 92, 179-191, (1997).

16. Raftery, A. E., **Bayes Factors and Bıc: Comment on a Critique of the Bayesian information Criterion for Model Selection**, Sociological Methods And Research, 27, 411-427, (1999).

17. Raftery, A. E., **Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models,** Biometrika, 83, 251-266, (1996).

18. Raftery, A.E., **Bayesian Model Selection in Social Research**, Sociological Methodology, Marsden, P.V. Cambridge, Mass., Blackwells, 111-196, (1995).

19. Bernard, G. A., **New Methods of Quality Control**, J. Roy. Statist. Soc. Ser. A, 226-255, (1963).

20. Roberts, H. V., **Probabilistic Prediction of a Model**, Ann. Statist., 6, 50-62, (1965).

21. Volınsky, C. T., Madıgan, D., Raftery, A. E. and Kronmal,. R. A., **Bayesian Model Avaraging in Proportional Hazard Models: Assesing the Risk of Stroke**, J. Roy. Statist. Soc. Ser. C, 46, 433-448, (1997).