


5-1-2004

Accurate Binary Decisions For Assessing Coronary Artery Disease

Mehmet Ali Cengiz

University of Ondokuz Mayıs, Samsun, Turkey, macengiz@omu.edu.tr

Follow this and additional works at: <http://digitalcommons.wayne.edu/jmasm>

 Part of the [Applied Statistics Commons](#), [Social and Behavioral Sciences Commons](#), and the [Statistical Theory Commons](#)

Recommended Citation

Cengiz, Mehmet Ali (2004) "Accurate Binary Decisions For Assessing Coronary Artery Disease," *Journal of Modern Applied Statistical Methods*: Vol. 3 : Iss. 1 , Article 16.

DOI: [10.22237/jmasm/1083370560](https://doi.org/10.22237/jmasm/1083370560)

Available at: <http://digitalcommons.wayne.edu/jmasm/vol3/iss1/16>

This Regular Article is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized editor of DigitalCommons@WayneState.

Accurate Binary Decisions For Assessing Coronary Artery Disease

Mehmet Ali Cengiz
University of Ondokuz Mayıs
Samsun, Turkey

Generalized linear models offer convenient and highly applicable tools for modeling and predicting the behavior of random variables in terms of observable factors and covariates. This paper investigates applications of a special case of generalized linear model to improve the accuracy of predictions and decisions adopting Bayesian methods, in the specific context of assessing coronary artery disease. The basic model is developed for this application using binary response. The results clearly demonstrate the potential advantages offered by this approach.

Key words: Bayesian methods, coronary artery disease

Introduction

The aim of this paper is to determine the probability of using Bayesian inference, in place of Classical inference, and to compare these two approaches and then to present new approach in assessing the probability of presence of Coronary artery disease. Multiple logistic regression was used, which is a special case of generalized linear models. This model is commonly used when the independent variables include both numerical and nominal measures and the outcome variable is binary, or dichotomous, having only two values. It requires no assumptions about the distribution of the independent variables.

Another advantage is that the regression coefficient can be interpreted in terms of relative risk in cohort studies or odds ratios in case-control studies. The Bayesian inference is based on the famous published posthumously by the Rev. Thomas Bayes in 1763. In this inference the numerical values allotted to probabilities do not relate to long-run frequencies and an attempt

is made to account for prior knowledge by quantitative measurement. The process of inference requires the evaluation of further integrals and the selection of appropriate prior.

In this paper a suitable prior distributions is presented. In some practical applications there is very little prior information available. In this case, the standard choice over recent years has been the invariant prior proposed by Jeffreys (1939). The other suitable priors may be Uniform, which is described many authors such as Bernardo and Smith (1994) and O'Hagan (1994).

The evaluation of integrals may be difficult analytically but numerical methods can overcome this difficulty. Dunsmore (1976) considered an asymptotic Bayesian approach to prediction analysis. Percy (1993) used this approach in the context of generalized linear models. Tierney and Kadane (1989) introduced The Laplace approximation that can be used to obtain a marginal of the posterior distribution. The above mentioned approaches were used and modified to binary data. By analyzing a set of data relating a real surgical problem (diagnosis of Coronary artery disease), several questions and suggestions arise regarding this application.

Mehmet Ali Cengiz is an Assistant Professor of Statistics at Ondokuz Mayıs University, Department of Statistics in Turkey. He graduated from the University of Salford in England with a Ph.D. in Applied Statistics in 1999. E-mail him at: macengiz@omu.edu.tr.

Coronary Artery Disease

Balcı et al (2000) previously investigated this surgical application. Their aim was to investigate the relationship between plasma insulin levels and the angiographical severity of coronary artery disease in male patients with normal glucose tolerance and unstable angina. The current work uses their data and results. Start by briefly reviewing the medical details that are relevant to the present analysis. Coronary Artery Disease is a progressive disease process that generally begins in childhood and has clinical manifestations in the middle to late adulthood.

Two decades ago, Coronary Artery Disease was considered to be a degenerative process because of the accumulation of lipid and necrotic debris in the advanced lesions. It is now recognized that it is a multifactorial process, which, if it leads to clinical sequelae, requires extensive proliferation of smooth muscle cells within the intima of the affected artery. The form and content of the advanced lesions of Coronary Artery Disease demonstrates the results of three fundamental biological processes.

These are: (1) proliferation of intimal smooth muscle cells, together with variable numbers of accumulated macrophages and T-lymphocytes; (2) formation by the proliferated smooth muscle of large amounts of connective tissue matrix, including collagen and elastic fibers (3) accumulation of lipid, principally in the form of cholesteryl esters and free cholesterol within the cells as well as in the surrounding connective tissues. The development of the concept of risk factors and their relationships to the incidence of coronary Artery Disease evolved from prospective epidemiological studies. These studies demonstrated a consistent association among characteristics observed at one point in time in apparently healthy individuals with the subsequent incidence of coronary artery disease in these individuals (Braunwald, 1992).

These associations include an increase in the concentration of plasma cholesterol, the incidence of cigarette smoking, hypertension, clinical diabetes, insulin levels, obesity, age or male sex, and occurrence of coronary artery disease. As a result of these associations, each

characteristic has been termed a risk factor and this terminology has been generally accepted and has become part of the scientific literature associated with this problem. The aim here is to develop a generalized linear model to calibrate coronary arterial stenoses against some risk factors, so that disease severity can be assessed with using some risk factors.

Bayesian Inference of Logistic Model to Binomial Data

Assuming n binomial observations of the form y_i , $i=1, \dots, n$ where $E(y_i) = p_i$ and p_i is the success probability corresponding to the i th observation, the linear logistic model for the dependence p_i on the values of the k explanatory variables $x_{1i}, x_{2i}, \dots, x_{ki}$, associated with that observation, is

$$\begin{aligned} \text{logit}(p_i) &= \log(p_i / (1 - p_i)) \\ &= \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} \end{aligned} \quad (1)$$

In order to fit a linear logistic model to a given set of data, unknown parameters must be estimated first. In Classical approach, these parameters are estimated using the methods of maximum likelihood. The likelihood function is given by

$$L(\beta_i; y_i) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1 - y_i} \quad (2)$$

The problem is to obtain estimations of parameters, which maximise the

$$l(\beta_i; y_i) = \sum_{i=1}^n y_i \log(p_i) + (1 - y_i) \log(1 - p_i).$$

Bayesian inference is used to obtain parameter estimations.

Assuming some training data $D = \{(Z_i, y_i); i = 1, \dots, n\}$ which consist of observed response vectors y_i and matrices of explanatory variables Z_i , typically one will observe Z_{n+1} for a new individual, and our aim is to predict the response vector y_{n+1} . The

conditional distribution of y_i given Z_i is assumed known as a function of unknown parameters contained in a vector β . The posterior predictive distribution of y_{n+1} given Z_{n+1} and the data D is given by

$$f(y_{n+1} | Z_{n+1}, D) = \int_{\beta} f(y_{n+1} | Z_{n+1}, \beta) f(\beta | D) d\beta \quad (3)$$

In the third equation (3), $f(\beta | D) \propto L(\beta; D) \times f(\beta)$ where L is the likelihood and f represents the prior density. The likelihood function is given in equation (2). If information about parameters before observing the data is vague, the use of the uniform prior distribution for a location parameter is supported by several researchers. The task of finding logically consistent realistic representations of prior ignorance meets some difficulties. In particular the uniform distribution may not represent ignorance. Jefferys (1967) proposed a solution using Fisher information matrix. There are different Jeffreys prior to the binomial experiments, so that posterior inference using the Jeffreys prior will violate the Likelihood Principle. So uniform and Jeffreys prior distributions for our application are used.

The required integrations in equation (3) and (5) are not feasible analytically and approximation methods are needed. Dunsmore (1976) considered an asymptotic Bayesian approach to prediction analysis. If we expand $f(y_{n+1} | Z_{n+1}, \beta)$ in equation (3) about the maximum likelihood estimate of β by Taylor's theorem, A first order approximation and second order approximation to the predictive distribution are then obtained by truncating the expanded series. The following equation for first order approximation is obtained.

$$f(y_{n+1} | Z_{n+1}, D) \approx f(y_{n+1} | Z_{n+1}, \hat{\beta}) \quad (4)$$

The Laplace approximation is useful for evaluating the multiple integral in equation (5) to predict disease severity, since the information matrix can be obtained without a lot of effort. The equation may be re-expressed (3) as

$$f(y_{n+1} | Z_{n+1}, D) = \frac{\int_{\beta} f(y_{n+1} | Z_{n+1}, \beta) L(\beta; D) f(\beta) d\beta}{\int_{\beta} L(\beta; D) f(\beta) d\beta} \quad (5)$$

From equation (5), the posterior expectation of $f(y_{n+1} | Z_{n+1}, D)$ can be expressed as the ratio

$$E\{f(y_{n+1} | Z_{n+1}, D)\} = \frac{\int_{\beta} f(y_{n+1} | Z_{n+1}, \beta) L(\beta; D) f(\beta) d\beta}{\int_{\beta} L(\beta; D) f(\beta) d\beta} \quad (6).$$

Referring to Tierney and Kadane (1986), it may be written

$$E\{g(\beta)\} \approx \left(\frac{\det \tilde{\Sigma}}{\det \Sigma} \right)^{1/2} \exp \left[n \left\{ \tilde{l}(\tilde{\beta}) - l(\hat{\beta}) \right\} \right]$$

where $\tilde{\beta}$ and $\hat{\beta}$ maximize $\tilde{l}(\beta) = (\log g + \log f + \log L) / n$ and $l(\beta) = (\log f + \log L) / n$, respectively, and $\tilde{\Sigma}$ and Σ are minus the inverse Hessians of $\tilde{l}(\beta)$ and $l(\beta)$ evaluated at $\tilde{\beta}$ and $\hat{\beta}$, respectively and n is the sample size for which data have been observed.

Methodology & Results

The data for the analyses were collected in 1996 – 1997 and presented in Table 1, at University Hospital in Erzurum, Turkey. One hundred consecutive men undergoing elective coronary angiography formed the study population. Eligible patients met the following criteria: (1) no history of diabetes; (2) normal fasting blood glucose; (3) no treatment with lipid lowering drugs (4) no antecedent history of myocardial infarction, coronary bypass, or angioplasty. Cardiovascular medications including β blockers, calcium antagonists, nitrates, aspirin, angiotensin-converting enzyme inhibitors were not discontinued before the study. A standard oral glucose tolerance test was performed 3 days before coronary angiography. Selective coronary

angiography was performed by Judkins technique in the right and left oblique views. 3 observers unaware of the laboratory results examined angiograms.

The luminal percent diameter narrowing was estimated by a consensus of the observers or by the mean of different measurements. Diameters stenosis $\geq 50\%$ were considered significant and these patients (68 patients) were assigned to the diseased one. Stepwise logistic regression was performed to evaluate the independence of risk factor effects on presence of Coronary Artery Disease.

For patient $i = 1, \dots, n$ expert judgments were used to classify each patient as healthy ($y_i = 0$) or diseased ($y_i = 1$) as mentioned above. After performing stepwise regression, Patient i has also has three covariates:

x_{1i} : age for patient i

x_{2i} : Log fasting insulin level for patient i x_{3i} :

Log Lp(a) (Lp(a): Lipoprotein (a))

Now consider the following model with using different prior distributions and different numerical approaches.

$$\begin{aligned} \text{logit}(p_i) &= \log(p_i / (1 - p_i)) \\ &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} \end{aligned}$$

where $E(y_i) = p_i$ and p_i is the success probability corresponding to the i th patient.

Considered is the same model as above with following cases.

1. case: the model above with Uniform prior and First Order approximation (corresponding to Classical approach using Likelihood method).

2. case: the model above with Uniform prior and Laplace approximation.

3. case: the model above with Jeffreys prior and First Order approximation.

4. case: the model above with Jeffreys prior and Laplace approximation.

The main aim for this section is to show how Bayesian inference in Bernoulli response models can be used to improve predictive accuracy in practice. Adopting a Bayesian approach to the analysis, a vague prior is used,

which is multiple uniform, and Jeffrey's prior because no specific prior information is available. Furthermore, we are merely demonstrating the potential of this model with different approaches in this paper; the goal is to develop a suitable informative prior in the future, to judge how sensitive the predictions are to the choice of prior.

Consequently, the joint posterior distribution, on which all predictive inference is based, is proportional to the likelihood function. In particular, the posterior predictive distribution for a new patient, with ages in vector x_1 , Log fasting insulin levels in vector x_2 and Log Lp(a)'s in vector x_3 , is

$$p(y | x_1, x_2, x_3, D) = \int_{\beta} p(y | x_1, x_2, x_3, \beta) f(\beta | D) d\beta \quad (7)$$

where $p(y | x_1, x_2, x_3, \beta)$ is the binomial sampling distribution defined by equation (1) and $f(\beta | D)$ is the joint posterior density, which is maximum likelihood function and prior distribution.

The assessment of diagnostic performance is now dealt with. Applied is the First order approximation and Laplace method using Fortran computer programs and subroutines from the NAG library to obtain approximate posterior predictive distributions as given by equation (7).

Two criteria to assess our predictive accuracy for each case were used. These are a binary loss function, corresponding to the percentage of correct classifications based on cross-validation of the training data set with a default classification threshold of 0.5 and the linear loss function

$$s_1 = \sum_{i=1}^{100} \{y_i(1 - \hat{p}_i) + (1 - y_i)\hat{p}_i\}$$

Table 1. Coronary Artery Disease

Patient i	y_i	x_{1i}	x_{2i}	x_{3i}	Patient i	y_i	x_{1i}	x_{2i}	x_{3i}
1	0	45	0.9542	1.6128	51	1	56	2.0334	2.0294
2	0	57	1.8261	1.3222	52	1	57	2.0792	1.7482
3	0	38	1.7404	0.4771	53	1	57	2.0128	1.2553
4	0	37	1.8129	0.3010	54	1	63	1.8921	0.9542
5	0	35	1.9031	1.2304	55	1	45	1.9031	1.7160
6	0	49	2.0934	1.9868	56	1	63	2.0000	1.2041
7	0	49	1.9912	1.0414	57	1	51	1.7634	1.0792
8	0	55	1.9243	1.4149	58	1	60	1.6812	1.5563
9	0	45	1.6021	0.8451	59	1	77	2.1461	2.2695
10	0	50	1.7634	1.2787	60	1	58	1.4771	2.4502
11	0	48	1.9031	1.0414	61	1	50	1.9031	1.6902
12	0	48	2.2355	1.2041	62	1	65	1.9031	2.0212
13	0	50	1.5315	1.5682	63	1	55	1.8062	1.9731
14	0	43	1.6021	0.7781	64	1	50	1.9031	2.0212
15	0	53	1.8325	1.7634	65	1	55	1.8325	1.5563
16	0	50	2.1703	1.6434	66	1	44	1.8195	0.4771
17	0	42	1.4472	1.3979	67	1	50	1.5798	2.4885
18	0	45	1.6021	0.9031	68	1	58	1.6021	0.9031
19	0	45	1.6021	1.6335	69	1	53	1.7482	2.5752
20	0	55	1.8129	1.9445	70	1	60	1.9243	0.6021
21	0	62	1.6021	0.8451	71	1	55	1.8062	2.4265
22	0	57	1.9031	1.2304	72	1	64	2.1461	1.8976
23	0	33	1.7404	0.6989	73	1	56	2.0492	1.4771
24	0	50	1.8129	1.3010	74	1	63	2.1399	1.5185
25	0	49	1.5051	1.5682	75	1	53	1.6532	0.0000
26	0	60	1.8195	1.7634	76	1	53	1.9345	1.3979
27	0	43	1.8976	0.3010	77	1	60	2.0086	1.3424
28	0	46	0.8451	1.5682	78	1	40	2.0253	1.6721
29	0	60	1.7781	0.6021	79	1	65	1.7781	0.4771
30	0	38	1.8261	1.3424	80	1	65	2.2878	1.2787
31	0	43	1.7324	1.2041	81	1	50	2.0792	1.6232
32	0	58	1.5563	0.7781	82	1	58	1.8195	1.1461
33	1	64	1.4771	1.2553	83	1	65	1.8808	1.5315
34	1	47	1.5315	1.6989	84	1	46	1.6812	1.6532
35	1	48	1.9777	1.6127	85	1	55	1.6021	2.4048
36	1	42	2.3617	0.4771	86	1	65	1.9138	1.5911
37	1	40	1.9445	1.9085	87	1	60	1.3424	1.3802
38	1	58	1.6021	1.1461	88	1	59	1.3424	2.0453
39	1	65	1.8062	2.1643	89	1	64	1.7634	1.6021
40	1	60	2.4232	1.4400	90	1	46	1.7324	1.2787
41	1	63	2.2041	0.0000	91	1	54	2.1367	1.5911
42	1	42	2.1732	1.5563	92	1	63	1.8808	1.6628
43	1	43	2.1614	1.1461	93	1	46	1.9031	2.3345
44	1	33	2.1987	1.2787	94	1	62	2.1987	1.2787
45	1	45	1.7243	1.2304	95	1	42	1.9138	1.8751
46	1	65	1.4771	1.4149	96	1	42	1.9031	1.4771
47	1	50	1.9542	2.2355	97	1	42	2.1461	2.4149
48	1	69	1.6989	1.59116	98	1	51	1.8808	1.9345
49	1	60	1.8451	1.7482	99	1	38	2.1004	1.6532
50	1	60	1.6532	1.9191	100	1	63	2.1367	1.4771

Table 2. Posterior predictive probabilities for the model with Laplace approximation and Jeffreys prior.

Patient i	\hat{p}_i	Patient i	\hat{p}_i	Patient i	\hat{p}_i	Patient i	\hat{p}_i
1	0.5885	26	0.0248	51	0.6241	76	0.4376
2	0.0817	27	0.4444	52	0.6095	77	0.5620
3	0.4042	28	0.6152	53	0.5062	78	0.3177
4	0.5689	29	0.1679	54	0.4961	79	0.3940
5	0.4168	30	0.3822	55	0.3533	80	0.6952
6	0.0303	31	0.3641	56	0.5775	81	0.4931
7	0.1665	32	0.2685	57	0.2610	82	0.4208
8	0.0635	33	0.3869	58	0.4703	83	0.6125
9	0.4532	34	0.2100	59	0.8436	84	0.2521
10	0.2116	35	0.4168	60	0.5185	85	0.5172
11	0.2220	36	0.2768	61	0.4316	86	0.6315
12	0.0737	37	0.3285	62	0.6848	87	0.2795
13	0.2737	38	0.3196	63	0.5217	88	0.3967
14	0.4954	39	0.6760	64	0.4949	89	0.5700
15	0.0716	40	0.6978	65	0.4558	90	0.2023
16	0.0333	41	0.4570	66	0.0890	91	0.5658
17	0.4794	42	0.3948	67	0.4463	92	0.6070
18	0.4436	43	0.3257	68	0.2722	93	0.4929
19	0.3159	44	0.2109	69	0.5749	94	0.6389
20	0.0435	45	0.1748	70	0.3984	95	0.3406
21	0.1728	46	0.4347	71	0.5974	96	0.2539
22	0.0701	47	0.5528	72	0.7214	97	0.5435
23	0.5950	48	0.6109	73	0.5434	98	0.4848
24	0.1847	49	0.5689	74	0.6651	99	0.3159
25	0.3045	50	0.5246	75	0.0854	100	0.6589

Table 3. Predictive accuracy results for the model with all cases.

Case	Binary Percentage	Linear loss
1	%79	0.3152
2	%81	0.2955
3	%81	0.2961
4	%83	0.2622

where $\hat{p}_i = P(Y_i = 1 | x_{1i}, x_{2i}, x_{3i}, D)$ from equation (7). Ultimately, the binary loss function is of most interest in diagnosing the disease, but the alternatives provide more insight into the predictive accuracy of the model with different approaches.

To illustrate the typical output from which loss functions are calculated, Table 2 presents the predictive probabilities for patients in the observed set of training data, based on the model with Jeffreys prior and Laplace approximation. The summary results for all cases investigated in Table 3.

First, is illustrated the improved predictive accuracy by adopting Bayesian inference here, over Classical approach. In case 1, uniform prior and First order approximation is used, which is the same as Classical approach (using the Likelihood function to obtain parameter estimations). Column 2 of Table 3 demonstrates this by presenting the percentage of diseased patients correctly diagnosed by each case, if costs are such that a threshold of 0.5 is appropriate. Note that, without further information, we could correctly diagnose 50 per cent of patients by chance alone, and that large values are desirable for the percentage of patients correctly diagnosed. Clearly, Bayesian approach with different priors and approximations performs consistently better than the classical approach.

Second, compared are the different priors and different approximations for the same model using two assessments criteria identified above: namely the binary and linear loss function. These results are presented in Table 3. Although large values are desirable for second column of these, small values are preferable for linear loss function. As expected, the model with Laplace approximation gives better results than the others.

Conclusion

This article described and discussed the properties and applications of multiple logistic regression models, suggesting simplifications and suitable approximations for a Bayesian analysis. Considered were different subjective priors, which are uniform, and Jeffreys, using different approximations, which are First order and Laplace approximation. It has also demonstrated how these prior distributions and approximations may be used and useful in an important application, relating to the diagnosis of coronary arterial disease.

References

- Balci, B. et al, (2000). The relationship between plasma insulin levels and angiographical severity of coronary artery disease. *Turk Kardiyol., Dern., Ars.* 28, 617-621.
- Bernardo, J. M. & Smith, A. F. M. (1994). *Bayesian theory*. John Wiley & Sons.
- Braunwald. E. (1992). *Heart disease*, Vol. 2. W. B. Saunders Company.
- Dunsmore, I. R. (1976). Asymptotic prediction analysis. *Biometrika*, 63, (3) 627-630.
- Jeffreys. H. (1939). *Theory of probability*. Oxford: Oxford University Press.
- O'Hagan. A. (1994). *Bayesian inference*. London: Edward Arnold.
- Percy. D. F. (1993) Prediction for generalized linear models. *Journal of Applied Statistics*, 20, 285-291.
- Tierney L., & Kadane J B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81, 82-86.