INFORMATION
SYSTEMS IN
MANAGEMENT

Information Systems in Management (2015) Vol. 4 (1) 53−63

THE ADVANTAGES OF BAYESIAN METHODS OVER CLASSICAL METHODS IN THE CONTEXT OF CREDIBLE INTERVALS

WIOLETTA GRZENDA

Institute of Statistics and Demography, Warsaw School of Economics (*SGH*)

The growing computational power of modern computer systems enables the efficient execution of algorithms. This is particularly important in Bayesian statistics, in which, nowadays, the key role is played by Markov Chain Monte Carlo methods. The primary objective of this work is to show the benefits arising from the use of Bayesian inference, especially confidence intervals in the context of logistic regression. The empirical analysis is based on "Household budgets" survey of Central Statistical Office. In this paper the unemployment among people over 55 will be investigated.

Keywords: Bayesian inference, confidence intervals, MCMC, logistic regression, unemployment

1. Introduction

Bayesian credible intervals, also known as credible regions (credible sets), play a similar role to classical credible intervals, but the philosophy of their composition and interpretation are quite different. The Bayesian approach gives the possibility of incorporating additional information that is external to the sample by prior distributions [8, 15]. This additional information may improve accuracy and credibility of estimations. The credible regions incorporate this prior information, while frequentist confidence intervals are based only on the sample data.

In frequentist methods, the parameters of a model are unknown, but fixed constants. Therefore, for a given sample, it cannot be specified whether the unknown value of this parameter is covered by this interval, or not. The randomness of this interval is based on the fact that for different samples we can obtain different realizations. The probability that the unknown parameter is within the confidence interval is either 0 or 1. However, in a long series of observed samples, the frequency of intervals, including true value of this parameter asymptotically equals $1-\alpha$ (for example [7]).

Moreover, the interpretation of classical confidence interval may be at times senseless; for example when only one sample exists and additional samples cannot be gathered. Then it is Bayesian approach that gives a reasonable interpretation and classical confidence interval should not be applied.

In Bayesian statistics, the estimated parameter is a random variable. Then the credible regions may be exactly estimated with a probability level given a priori. The Bayesian credible region estimated from the current sample contains the estimated parameter with the given probability. This interpretation seems intuitive and frequentist confidence intervals are thus misinterpreted as Bayesian intervals [3].

For small sample surveys, Bayesian analysis is more accurate as asymptotic approximation is not used [1]. However, it is worth mentioning that the impact of a priori distribution on a posterior distribution may be more significant.

The empirical examples presented in this paper refer to the analysis of unemployment among older people. There is no commonly accepted age over which an individual enters this age group [4]; usually these are employees aged 50 and older or aged 55 and older. In this paper, people aged 55 and older have been investigated as the retirement age has been raised recently. The chances of finding a job depend mainly on the situation on the labour market and some demographic and socio-economic characteristics of an individual. In unemployment studies, logit models have been most frequently used (for example [5], [12]). In this paper, the Bayesian logistic model has been used to analyse the impact of different characteristics on one's chances to find a job.

2. The definition of Bayesian confidence interval

Let Θ be the parameter space, C be subspace of this space, $C \subset \Theta$. Moreover let

$$p(\theta \mid \mathbf{x}) \propto p(\theta) p(\mathbf{x} \mid \theta) \qquad (1)$$

be posterior distribution of θ. θ can be a parameter vector, as will be the case in next part of this paper.

Then the probability that parameter θ is in the space C is given by

$$P(\theta \in C \mid \mathbf{x}) = \int_C p(\theta \mid \mathbf{x}) d\theta = 1 - \alpha,\ 0 < \alpha < 1. \tag{2}$$

This probability is the degree of belief that unknown parameter θ is in the space C, on condition that observations are available and we have a priori knowledge on this parameter. The region C is called Bayesian credible region. Moreover, there are many possible credible regions for a given probability level α, the smallest of them has been defined by formulas 3 and 4. It has been called the highest posterior density region (HPD).

If for the subset C^* the parameter space Θ, $C^* \subset \Theta$ holds

$$P(\theta \in C^* \mid \mathbf{x}) = 1 - \alpha,\ 0 < \alpha < 1 \tag{3}$$

and for every $\theta_1 \in C^*$ and $\theta_2 \notin C^*$ holds

$$p(\theta_1 \mid \mathbf{x}) \geq p(\theta_2 \mid \mathbf{x}), \tag{4}$$

then C^* is called $100(1-\alpha)\%$ the highest posterior density region.

The highest posterior density region is the region for which the minimum density of any point within that region is equal to or larger than the density of any point outside that region. This property refers only to Bayesian credible intervals and does not hold true in classical statistics [2, 10, 13].

3. Bayesian logistic regression

The logistic regression models explicate the relationship between a dependent binary or dichotomous variable and one or more independent variables [16]. In this paper, the binomial logistic regression will be investigated. Let $(X_1, \ldots, X_p)$ be a vector of independent variables, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ a vector of regression coefficients. Moreover, let π be the probability of success i.e. of obtaining by a dependent variable one of two possible values $y_i = 1$ or $y_i = 0$ ($i = 1, \ldots, n$), on condition that independent variables have given values.

Let

$$\operatorname{logit}(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p. \tag{5}$$

Then classical binomial regression model is given by

$$\pi = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p}}. \tag{6}$$

The likelihood function for a data set of n observations is

$$L(\boldsymbol{\beta};\mathbf{y})=\prod_{i=1}^{n}\left[\left(\frac{e^{\beta_0+\beta_1 X_{i1}+\ldots+\beta_p X_{ip}}}{1+e^{\beta_0+\beta_1 X_{i1}+\ldots+\beta_p X_{ip}}}\right)^{y_i}\left(1-\frac{e^{\beta_0+\beta_1 X_{i1}+\ldots+\beta_p X_{ip}}}{1+e^{\beta_0+\beta_1 X_{i1}+\ldots+\beta_p X_{ip}}}\right)^{(1-y_i)}\right] \tag{7}$$

In this paper the Bayesian approach to logistic regression has been investigated [1, 6, 8]. Bayesian method needs an appropriately defined prior distribution. In the case of regression models, for regression coefficients $\boldsymbol{\beta}$ we choose p–dimensional normal prior distribution $N_p(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$, where $\boldsymbol{\mu}_0$ denotes the prior mean vector, and $\boldsymbol{\Sigma}_0$ denotes the prior covariance matrix. For $j=1,\ldots,p$

$$\beta_j \sim N(\mu_j,\sigma_j^2). \tag{8}$$

Then the posteriori distribution by using Bayes' theorem is given by

$$\begin{aligned} p(\boldsymbol{\beta}\,|\,\mathbf{y})=&\prod_{i=1}^{n}\left[\left(\frac{e^{\beta_0+\beta_1 X_{i1}+\ldots+\beta_p X_{ip}}}{1+e^{\beta_0+\beta_1 X_{i1}+\ldots+\beta_p X_{ip}}}\right)^{y_i}\left(1-\frac{e^{\beta_0+\beta_1 X_{i1}+\ldots+\beta_p X_{ip}}}{1+e^{\beta_0+\beta_1 X_{i1}+\ldots+\beta_p X_{ip}}}\right)^{(1-y_i)}\right]\\ &\times\prod_{j=1}^{p}\frac{1}{\sqrt{2\pi}\sigma_j}\exp\left(-\frac{1}{2\sigma_j^2}(\beta_j-\mu_j)^2\right)\end{aligned} \tag{9}$$

In Bayesian approach, inference from any element of parameter vector $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)$ is held on the posterior marginal distribution. Then the distribution is obtained by integrating the remaining part of posterior distribution. The Bayesian confidence intervals for the elements of parameter vector $\boldsymbol{\beta}$ are calculated by formulae presented in the second part of this paper.

In practice, for models with a large number of parameters, the simulation methods are used to generate samples from an arbitrary posterior distribution. Currently Markov Chain Monte Carlo Methods (MCMC) are used in estimations [17]. The Markov Chain Monte Carlo Methods are based on ergodic Markov chain, which over time converges into the stationary distribution. In Bayesian statistics, this stationary distribution is called posterior distribution. In this paper ARMS algorithm (Adaptive Rejection Metropolis Sampling Algorithm) has been used [11].

4. Empirical example

In this paper, a data set from the survey of Central Statistical Office of Poland – “Household budgets in 2009” and “Household budgets in 2010” has been used. In the first survey, 37302 households including 108038 respondents have been examined and in the second one 37412 households with 107967 respondents. Professionally active individuals are people aged 15 and older, either employed or unemployed. In line with the aim of this research, only professionally active people aged 55 and older have been taken into consideration. Therefore, 4420 respondents have been chosen for analysis with the unemployment rate 5.75% for 2009 and 8.39% for 2010. By unemployed we mean people, who were looking for a job and were ready to take a job ‘this or the next week’. Farmers, gardeners, foresters and fishermen have not been investigated in this survey.

Table 1. The independent variables of models

Variable	Levels		Percent	
			2009	2010
Sex	man	1	65.63	64.22
	woman	2	34.37	35.78
Marital status	unmarried, separated or divorced, a widower, a widow	1	21.04	20.37
	married	2	78.96	79.63
Education status	higher	1	21.61	17.49
	post-secondary	2	2.87	6.47
	secondary professional	3	27.01	25.21
	secondary general	4	6.74	7.37
	basic vocational	5	29.68	31.37
	primary school	6	12.08	12.08
Region of Poland	central (łódzkie, mazowieckie)	1	23.87	23.45
	south (małopolskie, śląskie)	2	18.17	20.23
	east (lubelskie, podkarpackie, świętokrzyskie, podlaskie)	3	13.05	13.37
	northwest (wielkopolskie, zachodniopomorskie, lubuskie)	4	17.01	16.43
	southwest (dolnośląskie, opolskie)	5	12.35	11.82
	north (kujawsko-pomorskie, warmińsko-mazurskie, pomorskie)	6	15.54	14.70
Place of living	city of 100 thousand residents and more	1	37.49	35.21
	city below of 100 thousand residents	2	30.00	31.46
	country	3	32.51	33.33
Age group	60 years old and older	1	27.40	27.37
	less than 60 year	2	72.60	72.63

Source: own analysis of the data "Household budgets" 2009 and 2010

The most frequently reported factors related to unemployment are: sex, marital status, education status, age and kind of previous job [5]. For this survey, based on initial modelling, the following independent variables have been chosen: sex, marital status, education status, region of Poland (EUROSTAT) where a respondent lives, place of living and age. In table 1 the analysed characteristics of professionally active people have been presented.

Estimation and verification of all the models has been performed using SAS system. In the model, the number of burn-in iterations has been set to 2000 and the number of iterations after burn-in has been set to 10000.

In this paper, three models with non-informative and informative prior distributions have been estimated [14, 15]. The first model for data from 2009 has been estimated with non-informative prior distributions. The second model for data from 2010 has been estimated with non-informative prior distributions, too. The third model for data from 2010 has been estimated with informative prior information obtained as posterior information from model for data from 2009. Due to the change in the definition of levels for variable education in 2010 as compared to 2009, in the third model non-informative prior distributions have been used for this variable. The second model may be used as a reference model for the third one.

Before undertaking any inference from posterior distribution the convergence of generated Markov chains has been verified by the Geweke test [9], (see table 2). For all the parameters of investigated models no indication has been found that the Markov chains have not converged at any level of significance. The convergence of Markov chains has been confirmed by other tests and trace plots.

Table 2. The Geweke test

	Model 1		**Model 2**		**Model 3**	
Parameter	z	Pr > |z|	z	Pr > |z|	z	Pr > |z|
intercept	-1.5608	0.1186	0.1383	0.8900	0.9359	0.3493
sex1	0.5340	0.5933	0.0702	0.9440	-0.9284	0.3532
marital_status1	0.7778	0.4367	0.3511	0.7255	0.3913	0.6955
education1	-0.0130	0.9897	-0.3186	0.7500	-0.3132	0.7541
education2	-0.2544	0.7992	-0.1022	0.9186	1.0149	0.3102
education3	-0.0360	0.9713	-0.9752	0.3295	0.2080	0.8352
education4	0.1872	0.8515	-1.1732	0.2407	-0.5251	0.5995
education5	0.7346	0.4626	-1.4060	0.1597	-0.3220	0.7475
age_group1	1.2325	0.2178	0.5944	0.5523	0.6107	0.5414
region1	1.5252	0.1272	0.1809	0.8565	-0.7499	0.4533
region2	0.7866	0.4315	0.3710	0.7106	-1.3048	0.1920
region3	0.9525	0.3409	0.0646	0.9485	-0.1916	0.8480
region4	1.1261	0.2601	0.0439	0.9650	-0.4046	0.6858
region5	1.2234	0.2212	-0.2813	0.7785	0.4374	0.6619
place_of_living1	1.6682	0.0953	-0.0351	0.9720	1.4024	0.1608
place_of_living2	1.8743	0.0609	0.8223	0.4109	-1.3640	0.1726

Source: own analysis of the data "Household budgets" 2009 and 2010.

In Table 3, the 95% highest posterior density regions (HPD) have been given for all models. These highest posterior density regions may be interpreted as: there is a 95% chance that the unknown parameters are in these regions. For any parameters (for example: marital status) large region differences for the first and the second models have been obtained. This may indicate that information from only one sample is insufficient to investigate a given occurrence. As investigated models are based on data from two successive years, obtained results should be similar, because no significant changes in the labour market were observed in this period. Including informative prior from the previous year in the analysis has improved the accuracy of estimation. The lower range of credible regions has been obtained in the model with informative prior distributions (Model 3) compared to model with non-informative prior distributions (Model 2), (see Table 4).

Table 3. The 95% highest posterior density regions (HPD)

Parameter	**Model 1**		**Model 2**		**Model 3**	
intercept	2.1541	2.1814	1.5498	1.5711	1.3880	1.3955
sex1	-0.4235	-0.4068	-0.5783	-0.5648	-0.5094	-0.4997
marital_status1	-0.0954	-0.0782	-0.3947	-0.3815	-0.2596	-0.2584
education1	2.4528	2.4863	2.2231	2.2499	2.2252	2.2499
education 2	0.5851	0.6226	2.5278	2.5750	2.5568	2.6026
education 3	1.4139	1.4352	1.3003	1.3172	1.3382	1.3531
education 4	0.8189	0.8479	0.9101	0.9333	0.9384	0.9601
education 5	0.8180	0.8358	0.9436	0.9582	0.9792	0.9922
age_group1	0.7404	0.7601	0.3802	0.3944	0.5053	0.5079
region1	-0.0972	-0.0747	0.1604	0.1787	0.3734	0.3735
region2	0.1592	0.1838	0.3281	0.3472	0.3323	0.3336
region3	-0.2341	-0.2092	-0.1092	-0.0901	0.1502	0.1503
region4	0.0429	0.0675	0.4599	0.4802	0.4076	0.4193
region5	-0.1407	-0.1158	0.0963	0.1166	0.0872	0.0984
place_of_living1	-0.1644	-0.1457	0.1847	0.1997	0.1563	0.1565
place_of_living2	-0.3791	-0.3618	-0.0895	-0.0758	-0.1405	-0.1400

Source: own analysis of the data "Household budgets" 2009 and 2010

The measure of simulation accuracy is Monte Carlo standard error (MCSE). The values of Monte Carlo standard errors for the analysed model are presented in Table 5. These results explicitly indicate that the best estimations have been obtained for the third model with informative prior distribution.

Table 4. The length of the highest posterior density regions

Parameter	**Model 2**	**Model 3**
intercept	0.0213	0.0075
sex1	0.0135	0.0097
marital_status1	0.0132	0.0012
education1	0.0268	0.0247
education 2	0.0472	0.0458
education 3	0.0169	0.0149
education 4	0.0232	0.0217
education 5	0.0146	0.0130
age_group1	0.0142	0.0026
region1	0.0183	0.0001
region2	0.0191	0.0013
region3	0.0191	0.0001
region4	0.0203	0.0117
region5	0.0203	0.0112
place_of_living1	0.0150	0.0002
place_of_living2	0.0137	0.0005

Source: own analysis of the data "Household budgets" 2009 and 2010

Table 5. The values of Monte Carlo standard errors

	Model 1		**Model 2**		**Model 3**	
Parameter	MCSE	MCSE /SD	MCSE	MCSE /SD	MCSE	MCSE /SD
intercept	0.00031	0.0437	0.00023	0.0419	0.00005	0.0228
sex1	0.00013	0.0298	0.00010	0.0280	0.00005	0.0184
marital_status1	0.00008	0.0187	0.00006	0.0163	0	0.0100
education1	0.00014	0.0157	0.00012	0.0170	0.00008	0.0130
education 2	0.00015	0.0155	0.00014	0.0119	0.00012	0.0106
education 3	0.00012	0.0212	0.00010	0.0219	0.00006	0.0162
education 4	0.00014	0.0181	0.00011	0.0181	0.00007	0.0131
education 5	0.00011	0.0238	0.00009	0.0233	0.00006	0.0180
age_group1	0.00006	0.0120	0.00005	0.0124	0.00001	0.0100
region1	0.00017	0.0296	0.00012	0.0264	0	0.0100
region2	0.00016	0.0261	0.00012	0.0247	0	0.0100
region3	0.00017	0.0270	0.00012	0.0255	0	0.0100
region4	0.00017	0.0274	0.00013	0.0255	0.00004	0.0111
region5	0.00016	0.0258	0.00013	0.0241	0.00003	0.0110
place_of_living1	0.00011	0.0223	0.00008	0.0205	0.00001	0.0100
place_of_living2	0.00010	0.0226	0.00008	0.0213	0	0.0100

Source: own analysis of the data "Household budgets" 2009 and 2010

The interval estimation does not give information which values from credible intervals are the most probable, therefore empirical results have been supplemented by the point estimation values (see Table 6). In this paper, only the results for model 3 with informative prior distributions have been presented. Based on the highest probability density interval [3], all variables are statistically significant. The odds ratio has been calculated as well for a more detailed interpretation of the obtained results.

Table 6. Descriptive statistics of the posterior sample

Parameter	Mean	Standard deviation	Odds ratio
sex1	-0.5046	0.00245	0.604
marital_status1	-0.2592	0.00041	0.772
education1	2.2374	0.00638	9.369
education 2	2.5794	0.01170	13.189
education 3	1.3455	0.00381	3.840
education 4	0.9494	0.00556	2.584
education 5	0.9859	0.00333	2.680
age_group1	0.5062	0.00084	1.659
region1	0.3734	0.00004	1.453
region2	0.3332	0.00043	1.395
region3	0.1502	0.00004	1.162
region4	0.4132	0.00313	1.512
region5	0.0931	0.00306	1.098
place_of_living1	0.1564	0.00007	1.169
place_of_living2	-0.1403	0.00017	0.869

Source: own analysis of the data "Household budgets" 2010

The obtained values indicate that men have 39% less chance of having a job than women. For unmarried people the chance of having a job is about 23% less than for married people. The individuals who have an education level higher than primary, are more likely to have a job; 168% more likely for basic vocational, 158% more likely for secondary general, 284% more likely for secondary professional. Moreover, it is over thirteen times more probable for individuals who have post-secondary education to have a job than for people with the lowest education level. For people with higher education it is over nine times more probable.

People who live in other region of Poland than north have more chance of having a job. Moreover, people who live in the central region of Poland have the biggest chance of having a job as compared to people who live in north. As for the place of living, people who live in big cities are 16.9% more likely to have a job than people who live in the country and people who live in small cities are 13.1%

less likely compared to people who live in the country. People aged 60 and older have 65.9% more chance of having a job than people who are under 60.

5. Conclusion

In this paper Bayesian confidence intervals have been investigated in the context of logistic regression model. This approach entails significantly larger computational costs than classical methods due to a higher model complexity. The models have been estimated using Markov chain Monte Carlo methods with Gibbs sampling. In this work, benefits arising from the use of Bayesian approach to modelling, especially when confidence intervals are determined, have been shown.

The primary advantage of the proposed approach is the ability to use out of the sample knowledge in the modelling process. This is particularly useful when modelling is performed on a regular basis as in the case of "Household budgets" research. The use of information from the preceding year yielded better parameter estimation. In particular, compared to the use of non-informative prior distributions, lower values of Monte Carlo standard errors have been attained.

Bayesian confidence regions yielded information on the range of change of estimated parameters with probability of 0.95. Moreover, this result was obtained based on this particular sample, while using a priori information from the previous survey. The use of informative prior distribution resulted in the significant reduction of highest posterior density region range as compared to the model using non-informative prior distribution.

Based on the estimation made, the feature that influences most the employment status of the persons 55 years old and older is education status. The results for the sex variable are different from the results of other studies, conducted for the entire population, for example from BAEL research results [18].

REFERENCES

[1] Albert J.H., Chib S. (1993) *Bayesian analysis of binary and polychotomos response data*, Journal of the American Statistical Association 88, 669-679.

[2] Bernardo J., Smith A. (2004) *Bayesian Theory*, Wiley Series in Probability and Statistics, Wiley & Sons, New York.

[3] Bolstad W.M. (2007) *Introduction to Bayesian statistics*, Wiley & Sons, USA.

[4] Błędowski P. (2013) *Older People in the labour market*, in: *Labour market and demographic change* (ed. Kiełkowska M.), Demographic Publications, 52-63 (in Polish).

[5] Collier W. (2003) *The Impact of Demographic and Individual Heterogeneity on Unemployment Duration: A Regional Study*, Studies in Economics, 0302.

[6] Congdon P. (2007) *Bayesian Statistical Modelling*, Wiley & Sons, New York.

[7] Fisz M. (1967) *Theory of Probability and mathematical statistic*, PWN, Warsaw (in Polish).

[8] Gelman A., Carlin J.B., Stern H.S., Rubin D.B. (2000) *Bayesian data analysis*, Chapman & Hall/CRC, USA.

[9] Geweke J. (1992) *Evaluating the Accuracy of Sampling-based Approaches to Calculating Posterior Moments*, in: Bernardo, J., Berger, J., Dawiv, A., Smith, A., Bayesian Statistics 4, 169-193.

[10] Gill J. (2008) *Bayesian method: a social and behavioral sciences approach*, Chapman & Hall/CRC, London.

[11] Gilks W., Best N., Tan K. (1995) *Adaptive rejection Metropolis sampling with Gibbs sampling*, Applied Statistics 44, 455-472.

[12] Grzenda W. (2011) *The use of decision trees and logistic regression models to analyse demographic and socio-economic factors influencing the chances of finding a job*, Economics Studies 95, 271-277 (in Polish).

[13] Grzenda W. (2012) *Introduction to Bayesian Statistics*, SGH Publishing House, Warsaw (in Polish).

[14] W. Grzenda. (2013) The significance of prior information in Bayesian parametric survival models, Acta Universitatis Lodziensis, Folia Oeconomica, 285, 31-39.

[15] Ibrahim J.G., Chen M.H. (2000) *Power Prior Distributions for Regression Models*, Statistical Science, 15 (1), 46-60.

[16] Larose D.T. (2006) Data Mining Methods and Models, Wiley, New York,USA.

[17] Robert Ch.P., Casella G. (2004) *Monte Carlo Statistical Methods*, Springer, USA.

[18] http://stat.gov.pl/cps/rde/xbcr/gus/pw_kwart_inf_aktywnosc_ekonomiczna_ludnosci_1-4kw2010.zip (in Polish, 2010.01.10).