EGE AKADEMİK BAKIŞ / *EGE ACADEMIC REVIEW*

Cilt 21 • Sayı 3 • Temmuz 2021
SS. 263/272
Doi: 10.21121/eab.960840
Başvuru Tarihi: 21.01.2021 • Kabul Tarihi: 23.06.2021

Comparison of Different Estimation Approaches in Rare Events Data

N.Ece BACAKSIZ[1], Selçuk KOÇ[2]

ABSTRACT

In social science researches, there may be cases where a category of the dependent variable is seen hundred times less (more) than the other category. Events like wars, mass migrations or coups in social sciences; an event of interest in binary variable(s) may have very low prevalence, resulting in low or even zero cell counts in one or two cells in the 2X2 tables of two factors. In this case, independent variable predict the dependent variable perfectly or almost perfectly, and this leads to an issue called complete or quasi-complete separation problem in statistical modelling. This study aims to compare three methods suggested in the literature for the quasi-complete separation in a real small dataset; penalized maximum likelihood (Firth-type), exact logistic regression and bayesian logistic regression. Methods were compared via odds ratios, odds' standard error estimates, confidence intervals and statistical significance. Parameter estimates were obtained under three different models with binary and continuous variables. Results show that all methods can provide convergence in the presence of quasi-complete separation. In conclusion, bayesian logistic regression estimates tend to be superior than the other methods in terms of estimation of standard errors.

Keywords: Rare events, zero cell count, quasi-complete separation, bayesian logistic regression, penalized maximum likelihood

1. Introduction

Frequency of occurence of a category of binary dependent variable may be considerably rare than the other, in medical, social and political science researches. For example a side effect of a drug may be seen only in out of 1000 patients; similar rare event cases may be observed in wars, coups, mass migrations as examples of this rarity in economic analysis (King and Zeng, 2001: 693). The level of rarity depends on the event's prevalence. However, in rare events, depending of the level of rarity and due to randomness, the existence of zero frequency cells can be seen in 2X2 tables between a dependent variable and an independent variable. Suppose that of the 28 patients who have similar complaints treated at the same psychiatry clinic, 15 get treatment by medication and 13 by psychotherapy. In table 1a and 1b, we show two examples of rare event scenarios we may end up observing from this patient population response distribution for these patients by therapy:

Table 1: Representative 2x2 tables for zero cell count

1.a Two zero cells			
Yes		Response	
		No	
Treatment type	Medication	0	15
	Psychotheraphy	13	0

1.b One zero cell			
Yes		Response	
		No	
Treatment type	Medication	0	15
	Psychotheraphy	10	3

[1]Dr., ecebacaksiz@gmail.com
[2]Prof.Dr., Kocaeli Üniversitesi İİBF İktisat Bölümü, selcukkoc@kocaeli.edu.tr

It is seen that in both cases, the therapy covers the entire part of the zero outcomes of the response to treatment. These situations occur if the responses and non-responses can be perfectly separated by a single risk factor or linear combination of risk factors and called separation (Heinze and Schemper, 2002: 2409). The first statement in Table 1.a (two zero cells) refers to complete separation and Table 1.b refers to quasi-complete separation. In both types, maximum likelihood estimates of logistic regression does not exist. In other words concavity can not be achieved in iterations leading infinite function and it is called as convergence failure. Convergence does not occur because one or more parameters in the model become theoretically infinite (Webb et al., 2004: 274).

Even in some cases where maximum likelihood estimates exist, one can suspect about the uniqueness of the maximum likelihood estimation and the reliability of the estimation results. Large odds ratio estimates, large standard error estimates and hence, wide confidence intervals put the models far from interprability. Separation is more common in small samples (Heinze and Schemper, 2002:2409). This is because the number of observations per cell is less and therefore probability of zero cells is higher.

This study aims to compare three methods suggested in the literature for the quasi-complete separation problem in a small dataset; penalized maximum likelihood (Firth type), exact logistic regression and bayesian logistic regression. For this aim, data obtained from 53 feasibility reports of public transportation investments subordinated by the Ministry of Transportation and Infrastructure of Republic of Turkey is used. The paper is organized as follows. Section 2 discusses the methods proposed for quasi-complete separation. Section 3 presents the literature review of empirical studies directly related to this study, followed by Section 4 providing the empirical analysis. In Section 5, discussions, conclusions and future work are provided.

2. Solutions for quasi-complete separation

There are some methods in the literature for the solution of the quasi-complete separation problem. Solutions like increasing the amount of data by artificially inflating the zero cells (https://support.minitab.com), deletion of problem variables, combining dummy variable categories can be useful in first step (Allison, 2008: 7). While these approaches may sound practicle to a degree, they either does not solve the problem at hand or they do so with a potential bias in estimation of odds ratios. Therefore, more appropriate approaches such as penalized maximum likelihood (Firth, 1993), exact logistic regression (Mehta and Patel, 1995) and bayesian logistic regression have been proposed. We now describe each of these methods briefly.

2.1. Penalized maximum likelihood

Logistic regression model is defined as $Prob(y_i = 1|x_i, \beta) = \pi_i = \frac{1}{\{1+\exp(-x_i\beta_r)\}}$ where $i = 1,2,,\dots,n$ denotes the sample of n observations and $r = 1,2,\dots,k$ independent variables. Maximum likelihood estimate of regression parameters are obtained as a result of score equation $\frac{\delta logL}{\delta \beta_r} = U(\beta_r) = \sum_{i=1}^{n}(y_i - \pi_i)x_{ir} = 0$ where L is likelihood function. Firth (1993) suggested to maximize $logL(\beta)^* = logL(\beta) + 1/2\log|I(\beta)|$ in order to obtain finite estimates and reduce bias where $1/2\log|I(\beta)|$ is penalty function. If this modification is applied to logistic model, the score equation becomes 'modified' score equation $U(\beta)^* = \sum_{i=1}^{n}\left\{y_i - \pi_i + h_i\left(\frac{1}{2} - \pi_i\right)\right\}x_{ir} = 0, \quad (r = 1,2,\dots,k)$. Here, $h_i's$ are the *i*th diagonal elements of the hat matrix $H = W^{1/2}X(X^TWX)^{1-}X^TW^{1/2}$ with nxp dimensions and $W = diag\left\{\left[\frac{1}{1+\exp(-X_i\beta)}\right]\left[1 - \frac{1}{1+\exp(-X_i\beta)}\right]\right\} \equiv diag[\pi_i(1-\pi_i)]$. After this, Firth-type estimates can be obtained iteratively the usual way until convergence is obtained:

$$\beta^{(s+1)} = \beta^{(s)} + I^{-1}(\beta^{(s)})U(\beta^{(s)})^*$$

Where the superscript (s) refers the s^{th} iteration (Henze and Schemper, 2002: 2411-2412; Eyduran, 2008: 326).

2.2. Exact logistic regression

The general idea was to base inferences on exact permutational distributions of the sufficient statistics that correspond to the regression parameters of interest, conditional on fixing the sufficient statistics of the remaining parameters at their observed values (King and Ryan, 2002: 164). The aim of exact conditional analysis is to determine how likely the observed response is with respect to all 2^n possible responses (Derr, 2009). To perform conditional inference the sufficient statistics for the β_j in the unconditional likelihood function are calculated as $T_j = \sum_{i=1}^{n} y_i x_{ji}$. The probability density function for $T = (T_0, T_1, \dots, T_p)'$, the vector of sufficient statistics, all binary sequences *y* that generate observable *t*.

$$P(T = t) = \frac{C(t)\exp(t'\beta)}{\prod_{i=1}^{n}[1 + \exp(x_i'\beta)]}$$

where $C(t) = \left|\left|\{y: y'x = t\}\right|\right|$ is the number of sequences generating t. If $\beta_o = (\beta_1, \dots, \beta_p)'$ is accepted as nuisance parameter, then the corresponding sufficient statistics for given β_o is T_0. Similarly, T_1, t_1 and X_1 is defined as parameters of interest. For creating conditional likelihood, nuisance parameters can be removed from the analysis by conditioning on their sufficient statistics:

$$P(T_p = t_p | T_0 = t_0) = \frac{P(T = t)}{P(T_0 = t_0)}$$

$$= \frac{C(t)\exp(t_p'\beta_p)}{\sum_u C(u, t_0)\exp(u'\beta_p)}$$

where $C(u, t_0)$ is the number of vectors such that $y'X_1 = u$ and $y'X_0 = t_0$. Exact logistic regression estimates of β_p is the value that maximizes the conditional likelihood (Derr, 2009; King and Ryan, 2002: 164).

Because the conditional distributions of sufficient statistics requires summing over discrete patterns of covariate values, relatively sparse and/or small data in particular patterns of categorical covariates often lead to degenerate estimates. The inclusion of continuous covariates only magnifies this issue of sparseness. Hence, this makes the exact method less attractive for researchers whose analysis includes combination of continuous and categorical independent variables (Zorn, 2005: 162).

2.3. Bayesian logistic regression

Bayesian analysis has following steps: it starts with a prior distribution on the unknown parameters and updates this with the likelihood of the data, yielding a posterior distribution which is used for inferences and predictions.

In binary logistic regression, likelihood function is written below for $y_i = 0\ or\ y_i = 1$:

$$p(y|\beta, X) = \prod_{i=1}^{n} \begin{Bmatrix} logit^{-1}(X_i\beta)\ if & y_i = 1 \\ 1 - logit^{-1}(X_i\beta)\ if & y_i = 0 \end{Bmatrix}$$

or

$$p(y|\beta, X) = \prod_{i=1}^{n} (logit^{-1}(X_i\beta))^{y_i}(1 - logit^{-1}(X_i\beta))^{1-y_i}$$

and the likelihood contribution from the i*th* subject is,

$$likelihood_i = \left(\frac{e^{\beta_0+\beta_1X_{i1}+\cdots+\beta_pX_{ip}}}{1+e^{\beta_0+\beta_1X_{i1}+\cdots+\beta_pX_{ip}}}\right)^{y_i}\left(1-\frac{e^{\beta_0+\beta_1X_{i1}+\cdots+\beta_pX_{ip}}}{1+e^{\beta_0+\beta_1X_{i1}+\cdots+\beta_pX_{ip}}}\right)^{1-y_i}$$

so the likelihood function over a data set of *n* subjects is then;

$$p(y|\beta, X) = \prod_{i=1}^{n}\left[\left(\frac{e^{\beta_0+\beta_1X_{i1}+\cdots+\beta_pX_{ip}}}{1+e^{\beta_0+\beta_1X_{i1}+\cdots+\beta_pX_{ip}}}\right)^{y_i}\left(1-\frac{e^{\beta_0+\beta_1X_{i1}+\cdots+\beta_pX_{ip}}}{1+e^{\beta_0+\beta_1X_{i1}+\cdots+\beta_pX_{ip}}}\right)^{1-y_i}\right]$$

A priori distribution of a parameter is a probability distribution that includes information that is not explicit about the parameter before analyzing the data. Existing literature proposes fully informative or noninformative priors in bayesian analysis. The issue of priors for logistic regression models can be found in Gelman et al. (2008) and Gelman et al. (2013). In this study, normal distribution (with 0 mean and 1 variance) was preferred considered to be most competible with data (with most plausible odds ratios and standard errors) similar to Greenland et al. (2000), Soliman et al. (2013) and Kocak (2017).

In third step, the likelihood function is updated with a priori distribution, and parameter estimates are obtained from this posterior distribution:

$$p(\beta_i|Y) = \frac{L(\beta_i|Y)p(\beta_i)}{\int L(\beta_i|Y)p(\beta_i)d\beta_i}$$

$L(\beta_i|Y)$ is likelihood of β_i and $p(y)$ is an integral. This statement transforms the maximum likelihood estimation into a sort of position of the posterior distribution. To summarize, the Bayesian approach refers to how to update existing knowledge with new knowledge (prior) (Gelman and Hill, 2007: 143; Rainey, 2016: 341; Cengiz, et al. 2013: 16).

3. Literature review

In this section, empirical studies which are related to the aim of the study directly were included.

Gavanji (2019) determines the factors effect occupational injuries in Saskatchewan . Between 2007-2016,

occupational injury rate in examined sample is 0,006%, which can be considered quitely rare. The study uses Firth logistic regression because conventional logistic regression including multiple categorical covariates. So, using Firth regression provides interpretable odss ratios and confidence intervals.

Kocak (2016), proposed an empirical bayesian estimation procedure for odds ratio of rare events and provides a formal test procedure. He compared the Bayesian approach with Exact Logistic and Firth approaches in terms of statistical power achieved through extensive simulations using no event and varying event rate scenarios. He showed that new method, which can be applied only for a single descriptive variable, is stronger in terms of keeping of Type 1 error and narrower confidence intervals compared to other methods. The method also predicted odds ratios even in extreme rare event scenarios.

Muchlinski et al. (2015) compared three types of logistic regressions including Firth type penalized regression with random forests method in rare events data. The dependent variable is a binary measure of whether a civil war onset occurred for a given country. Results show that the alghoritmic random forest approach provides significantly more accurate predictions on such a dependent variable than any of the logistic regression models implied.

Van Der Paal (2014) evaluated the performance of binomial linear regression model using different link functions and different primary distributions in rare events data. Four different data sets were obtained from UCI (Machine Learning Repository) database which have different rarity levels. Root mean squared error (RMSE), misclassification error rate (ER), the false positive rate (FPR), false negative rate (FNR) and the area under ROC curve (AUC) were used as performance measures. the relevant part of the study can be summarized as follows: at 1.34% rarity, except for the AUC, Bayes method with weakly informative prior distribution, gives better results than the Firth method. At 3.7% rarity, Bayes method shows better results than Firth's. At 14.3 % rarity, except for FPR and at 52.1% rarity except for ER and FPR, Bayes method gives better results than Firth's method.

Soliman, et al. (2013) studied small data sets which have separation problem. Firth bias-corrected regression, exact logistic regression, penalized logistic regression, removal of the variable causing separation, and a Bayesian logistic model with a weakly informative prior were compared in means of empirical performance. Estimation results show that The Bayesian model produced plausible confidence intervals of odds ratio. Firth's method produced implausibly large parameter estimates and wide confidence intervals.

Botes (2013) compared the Firth's method, exact logistic regression and hidden logistic regression under complete and quasi-complete separation for different data sizes. These three methods were compared by Pearson chi-square and Hosmer-Lemeshow test statistics. According to the results, exact logistic regression provides good estimation results when the common variables are the same data type, particularly categorical. In other words, degenerate results are obtained in combination of both categorical and continuous variables. Firth method achieved significant results when both categorical and continuous variable combinations were used. The hidden regression model performed well in terms of model significance, but in particular the non-significant coefficient estimates in the small sample. In general, the probabilities of each of the three methods in the large sample are very close to the observed values.

Guns and Vanacker (2012) has considered the rare events as natural hazards, in geomorphology field. They introduce some probabilistic approach based on Monte Carlo simulations in rare event logistic regression where the dichotomous dependent variable indicates the presence or absence of a landslide. They conclude that the modified rare event logistic regression based on Monte Carlo simulations to estimate the robustness of the regression estimates prevents instability of the results due to sampling bias, even in small data sets.

4. Empirical analysis

In this section, empirical analysis is shown including data, variables and results.

4.1. Data and variables

This study aims to compare three methods suggested in the literature for the quasi-complete separation problem in a small dataset; penalized maximum likelihood (Firth type), exact logistic regression and bayesian logistic regression. For this aim, data obtained from 53 feasibility reports[3] of public transportation investments generated between 2011-2015 years subordinated by

[3] In Turkey, investment feasibility report is requested by the Ministry of Transportation and Infrastructure for investments which have investment amount of 10 million TLs and above by 2018.

the Ministry of Transportation and Infrastructure of Republic of Turkey is used.

The comparison of methods was investigated under three different models. The first model only includes the problem variable, a qualitative variable is added in second model and a continuous variable is added in third model. In doing so, it is aimed to see if estimation performance of methods with distinct types of variable combinations would change. For the methodology and path, Kocak (2017) and Botes (2013) was taken as reference. Descriptive statistics are in the Appendix A. Variables used in this study are as follows:

Dependent variable

Cost/benefit ratio (C/B): It is a binary variable created by making the cost / benefit ratio calculated during economic analysis of public transport investments. It takes value of 1 if an investment's benefit is equivalent to or exceeds its costs, and 0 otherwise.

Independent variables

Sector: It shows the sector where investment is implemented. It takes value of 1 if the investment is applied on highway, and 0 otherwise (seaway or airway). This variable is also causes separation.

Region: It shows the region where investment is implemented. It is created by the electricity consumption of regions as a regional development indicator[4]. It takes value of 1 if the region (s) where the investments covers the 'top' by the electricity constumption classification, and 0 otherwise.

Implementation period: A continuous variable shows the estimated implementation period of an investment.

4.2. Empirical findings

The 2x2 table between the variable of the dependent variable and the sector variable is as follows:

Table: 2x2 table between C/B and sector

C/B	Sector		
	Highway	Other	Total
<1 (Coded as)	0	19	19
≥1 (Coded as 1)	23	11	34
Total	23	30	53

The presence of zero count in one cell is seen in Table 2 between the C/B and the highwaysector variable. Hence, quasi-complete separation problem exists. In this case, maximum likelihood estimates are not available. After detecting the quasi-complete separation in the data, Firth-type penalized maximum likelihood, exact logistic regression and bayesian logistic regression was applied. Estimations are separately made under three the models in which only the sector (problematic), sector and region (binary) variable, sector and duration of implementation (continuous) includes. Model statistics are in Appendix B. Estimation results are below:

As seen in Table 3, type penalized maximum likelihood estimates have the highest odds ratios and standard errors. The method also has big standard errors, the confidence intervals of odds are highest among other methods. Exact logistic regression method's odds ratios are smaller than Firth's and higher than bayesian method's. Nevertheless, exact method can not give interpretable standard error estimate for the problematic variable, which leads to an upper limit of confidence interval as +infinity. Bayesian logistic regression estimates seem more rational compared to other two: in the context of more plausible odds ratios, small standard errors and narrow confidence intervals. When odds' significances are examined, Firth-type estimates can be considered as having the smallest probability of significance- due to bigger standard error.

Table 3: Model with the problematic variable

Method	Variable	Odds Ratios	St.error	C.I. (%95)	Prob.
Penalized (Firth-type) maximum likelihood	Sector	79.69	117.69	(4.41, 1440.28)	0.003
Exact logistic regression	Sector	49.70	N/A	(7.51 +INF)	≤0.0001
Bayesian logistic regression	Sector	14.03	0.2970	(3.44, 41.66)	≤0.0001

[4] For the regional electricity constumption classification see General Directorate of Energy Affairs Bulletin (2015): http://www.enerji.gov.tr/File/?path=ROOT/1/Documents/E%C4%B0GM%20Periyodik%20Rapor/Mart-Nisan%20B%C3%BClteni_son.pdf

Tablo 4: Model with the problematic and a binary variables

Method	Variable	Odds	St.error	C.I. (%95)	Prob.
Penalized (Firth-type) maximum likelihood	Sector	76.31	116.62	(3.81, 1525.59)	0.005
	Region	0.12	0.10	(0.02, 0.68)	0.016
Exact logistic regression	Sector	43.37	N/A	(6.12, +INF)	≤0.0001
	Region	0.11	0.10	(0.009, 0.77)	0.021
Bayesian logistic regression	Sector	13.14	0.35	(3.37, 39.54)	≤0.0001
	Region	0.32	0.006	(0.09, 0.76)	0.019

It can be seen from Table 4 that estimation results do not differ much for problematic variable. In other words, adding a binary variable does not seem changing the results much on behalf of the problematic variable. Penalized maximum likelihood estimates have the biggest odds ratios, with smallest significance, also bigger standard errors with largest confidence intervals, Exact logistic regression, similar to the results above, does not have standard error estimates of problematic variable. Bayesian logistic regression has the most interpretable results again for the problematic variable. When we see estimation results for the variable added, odds ratios are closer in all methods. Standard error estimates are small, with the smallest estimated in bayesian logistic regression. Exact logistic regression has the estimation of standard error and upper limit of confidence intervals for added variable.

In Table 5, estimates for the model including problematic variable and the continuous variable. From the table, it can be seen that odds ratios are estimated smaller than the previous models for the problematic variable; penalized maximum likelihood estimated odds ratio is 60.03, exact odds ratio is 37.34 and bayesian odds is 10.61. Nevertheless, big standard error and large confidence intervals problems of Firth-type estimations exist. Similarly, exact logistic regression still does not have no standard error for problematic variable but has infinite upper limit of confidence interval. Bayesian logistic regression has still the most plausible results.

Tablo 5: Model with the problematic and a continuous variables

Method	Variable	Odds	St.error	C.I. (%95)	Prob.
Penalized (Firth-type) maximum likelihood	Sector	60.03	88.27	(3.36, 1071.26)	0.005
	Implementation period	1.62	0.69	(0.70, 3.76)	0.255
Exact logistic regression	Sector	37.34	N/A	(5.64, +INF)	≤0.0001
	Implementation period	1.70	0.75	(0.67, 4.76)	0.299
Bayesian logistic regression	Sector	10.61	0.32	(3.22, 38.06)	≤0.0001
	Implementation period	0.27	0.07	(0.09, 0.80)	0.223

5. Discussions

In some data sets, a category of binary variable / variables can be seen very low in total responses, even with zero cell counts in one or two cells in the 2X2 tables between the dependent and the independent. In this case, independent variable predict the dependent variable perfectly and separation problem arises. Separation, complete or quasi-complete, causes other problems as convergence failure of maximum likelihood, unstable odds ratios process, unreliable estimates (Kocak, 2017:2). In separation, there are solution methods which aim to eliminate zero cell from 2x2 tables like increasing the amount of data deletion of problem variables, combining dummy variable categories simply. But in some cases, these methods can not be implemented or fail. Other solution techniques that provide reliable maximum likelihood estimates are penalized maximum likelihood approach, exact logistic regression and bayesian logistic regression.

This study aims to compare different estimation approaches mentioned above when quasi-complete seperation exists in a rare event data. For this aim, data obtained from 53 feasibility reports of public transportation investments generated between 2011-2015 years subordinated by the Ministry of Transportation and Infrastructure of Republic of Turkey is used. During the statistical analysis of data, zero cell was seen in the 2x2 table of dependent and the sector variable. Then, the methods proposed in the literature have been tried to solve the problem of separation. Hence, logistic regression estimates could not be obtained. Then, estimation procedures in separation case was impemented. Analyzes performed with different variable combinations. All three methods achieved convergence. To summarize the results generally, the penalized maximum likelihood estimations result in very high odds ratios. Also the confidence intervals of odds ratios are much larger compared to the other methods'. The odds of exact logistic regression are smaller than penalized maximum likelihood method and larger than the bayesian method. In all three models, exact results cannot give standard error estimates, besides, the upper limit of confidence intervals of problematic variable go towards +Inf. The bayesian logistic regression provides reasonable odds ratios estimates, smaller standard errors and narrower confidence intervals compared to other two methods. Bayesian method also have similar and smaller p-values of odds ratios. The methods do not differ in superiority in three models created with different types of variables. For all that, when continuous variable included odds ratios' and standard errors estimates qualitatively smaller in penalized maximum likelihood approach. It is obvious that each method has its own advantages. Nevertheless, bayesian logistic regression estimates has a tendency to be superior compared to other methods in analyzing the data used in this study due to the relatively reasonable difference ratios estimates, small standard error estimates and narrower confidence intervals.

References

Allison P.D. (2008). Convergence failures in logistic regression. In: Proceedings of the SAS Global Forum 2008 Conference. SAS Institute Inc., Cary, NC. http://www2.sas.com/proceedings/forum2008/360-2008.pdf

Cengiz, M.A., Terzi, E. Şenel, T. ve Murat, N. (2013). Lojistik regresyonda parametre tahmininde Bayesci bir yaklaşım. *Afyon Kocatepe Üniversitesi Fen Bilimleri Dergisi*, 12(2012), 15-22.

Derr R.E. (2009). Performing exact logistic regression with the SAS System-Revised 2009. Proceedings of the Twenty-fifth Annual SAS Users Group International Conference; Cary, NC; 2009: Citeseer.

Devika, S. Jeyaseelan, L. ve Sebastian, G. (2016). Analysis of sparse data in logistic regression in medical research: a newer approach. *Journal of Postgraduate Medicine*, 62(1), 26-31.

Eyduran, E. (2008). Usage of penalized maximum likelihood estimation method in medical research: an alternative to maximum likelihood estimation method, *JRMS* 13(6), 325-330.

Firth D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika,* 80(1), 27-38.

Gavanji, R. (2019). *Penalized Regression Methods for Modelling Rare Events Data with Application to Occupational Injury Study* (Doctoral dissertation, University of Saskatchewan).

Gelman, A. ve Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, USA.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.V., Vehtari, A., Rubin, D.B. (2013). *Bayesian Data Analysis,* Third Edition. Chapman and Hall, London.

Gelman, A., Jakulin, A., Pittau, M.G., and Su, Y. (2009). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics,* 2(4), 1360–1383.

Greenland, S., Schwartzbaum, J.A., Finkle, W.D. (2000). Problems dur to small samples and sparse data in conditional regression analysis. *American Journal of Epidemiology*, 151(5), 531-539.

Guns, M., and Vanacker, V. (2012). Logistic regression applied to natural hazards: rare event logistic regression with replications. *Natural Hazards and Earth System Sciences*, 12(6), 1937-1947.

Heinze, G. And Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine,* 21, 2409-2419.

King, E.N. ve Ryan, T.P. (2002). A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression. *The American Statistician*, 56(3), 163-170.

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis.* 9(2), 137-163.

Kocak, M. (2017). An empirical Bayesian approach in estimating odds ratios for rare or zero events. *Turkiye Klinikleri J Biostat*, 9(1), 1-11.

Mehta, C.R. and Patel, N.R. (1995). Exact logistic regression: theory and examples. *Statistics in Medicine,* 14(19), 2143-2160.

Muchlinski, D., Siroky, D., He, J., and Kocher, M. (2016). Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis*, 87-103.

Paal, V.D. (2013). A comparison of different methods for modelling rare events data. Master Thesis. Universiteit Gent, Belgium.

Rainey, C. (2016). Dealing with separation in logistic regression models. *Political Analysis,* 2016(24), 339-355.

Soliman, A.M.A., MacLehose, R.F. and Carlson, A. (2013). Bayesian models with a weakly informative prior: A useful alternative for solving sparse data problems. *Value In Health.* 16(3), A48-A49.

Webb, M.C., Wilson, J.R. ve Chong, J. (2004). An analysis of quasi-complete binary data with logistic models: applications to alcohol abuse data. *Journal of Data Science,* 2(2004), 273-285.

Zorn, C. (2005). A solution to separation in binary logit models. *Political Analysis*, 13,157-170.

''What are complete separation and quasi-complete separation?". https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/supporting-topics/regression-models/what-are-complete-separation-and-quasi-complete-separation/ (04.09.2019).

"Bayesian inference for logistic regression parameters" www.medicine.mcgill.ca/epidemiology/.../bayeslogit.pdf (05.09.2019).

Appendix A: Descriptive statistics of variables

Variable	Frequency	Percentage
Cost/benefit ratio		
≥1	33	62.26
<1	20	37.74
Sector		
Highway	23	43.40
Otherwise	30	56.61
Region (according to electricity constumption)		
Upper region	20	37.74
Otherwise	33	67.26
Implementation period		
Minimum	Maximum	Mean
1	7	3.09434

Appendix B: Model statistics

Model with problematic variable		
Penalized maximum likelihood	Wald chi2	8.79
	Prob>chi	0.0030
	Penalized log-likelihood:	-19.60
	AIC	43.21
	BIC	47.15
Exact logistic regression	Model score	22.27
	Pr>score	≤0.0001
Bayesian logistic regression	Log-marginal likelihood	-27.0972
	Acceptance rate	0.26
	DIC	47.77
Model with problematic and a binary variable		
Penalized maximum likelihood	Wald chi2	11.62
	Prob>chi	0.003
	Penalized log-likelihood:	-15.87
	AIC	37.75
	BIC	43.66
Exact logistic regression	Model score	26.85
	Pr>score	≤0.0001
Bayesian logistic regression	Log-marginal likelihood	-24.94
	Acceptance rate	0.26
	DIC	42.63

Model with problematic and a continuous variable		
Penalized maximum likelihood	Wald chi2	8.76
	Prob>chi	0.01
	Penalized log-likelihood:	-17.97
	AIC	41.95
	BIC	47.86
Exact logistic regression	Model score	22.79
	Pr>score	≤0.0001
Bayesian logistic regression	Log-marginal likelihood	-29.92
	Acceptance rate	0.25
	DIC	47.89