mathematics

MDPI

Article

Use of Bayesian Markov Chain Monte Carlo Methods to Model Kuwait Medical Genetic Center Data: An Application to Down Syndrome and Mental Retardation

Reem Aljarallah [1] and Samer A Kharroubi [2,*]

[1] Department of Statistics and Operations Research, Kuwait University, Kuwait, P.O. Box 5969, Safat 13060, Kuwait; reema@kuc01.kuniv.edu.kw

[2] Department of Nutrition and Food Sciences, Faculty of Agricultural and Food Sciences, American University of Beirut, P.O. Box 11-0236, Riad El Solh 1107-2020, Beirut, Lebanon

[*] Correspondence: sk157@aub.edu.lb

Abstract: Logit, probit and complementary log-log models are the most widely used models when binary dependent variables are available. Conventionally, these models have been frequentists. This paper aims to demonstrate how such models can be implemented relatively quickly and easily from a Bayesian framework using Gibbs sampling Markov chain Monte Carlo simulation methods in WinBUGS. We focus on the modeling and prediction of Down syndrome (DS) and Mental retardation (MR) data from an observational study at Kuwait Medical Genetic Center over a 30-year time period between 1979 and 2009. Modeling algorithms were used in two distinct ways; firstly, using three different methods at the disease level, including logistic, probit and cloglog models, and, secondly, using bivariate logistic regression to study the association between the two diseases in question. The models are compared in terms of their predictive ability via R^2, adjusted R^2, root mean square error (RMSE) and Bayesian Deviance Information Criterion (DIC). In the univariate analysis, the logistic model performed best, with R^2 (0.1145), adjusted R^2 (0.114), RMSE (0.3074) and DIC (7435.98) for DS, and R^2 (0.0626), adjusted R^2 (0.0621), RMSE (0.4676) and DIC (23120) for MR. In the bivariate case, results revealed that 7 and 8 out of the 10 selected covariates were significantly associated with DS and MR respectively, whilst none were associated with the interaction between the two outcomes. Bayesian methods are more flexible in handling complex non-standard models as well as they allow model fit and complexity to be assessed straightforwardly for non-nested hierarchical models.

Keywords: Bayesian methods; Kuwait Medical Genetic Center; regression modeling; bivariate logistic regression; Markov chain Monte Carlo

check for updates

Citation: Aljarallah, R.; Kharroubi, S.A Use of Bayesian Markov Chain Monte Carlo Methods to Model Kuwait Medical Genetic Center Data: An Application to Down Syndrome and Mental Retardation. *Mathematics* **2021**, *9*, 248. https://doi.org/10.3390/math9030248

Academic Editors: Francisco-José Vázquez-Polo and Sorana D. Bolboacă

Received: 14 December 2020
Accepted: 23 January 2021
Published: 27 January 2021

1. Introduction

Logit, probit and complementary log-log or cloglog models are the most widely used members of the family of generalized linear models when binary dependent variables are available. Binary outcomes usually arise in many areas of applications in health sciences such as epidemiology and biomedical studies. These binary data are often multivariate or correlated, so it is of increasing interest to develop models that maintain marginal logistic interpretation pertaining to individual outcomes while taking into consideration the dependency structure.

Two common frequentist approaches have been proposed for the correlated binary data, namely marginal logistic regression via generalized estimation equations [1–4] and mixed effects logistic regression [5,6]. The former approach is often used as parameter interpretation is quite simple and there is robustness to misspecification of the correlation structure [7]. The second approach does not perform well in integrating out a random effects model from a mixed effects model, though this is not the case with the generalized estimation equations approach. Although likelihood-based approaches to marginal logistic fitting have also been proposed, the complexity of model specification is a considerable practical limitation [8,9].

When multivariate categorical data is being analyzed, Bayesian approaches have shown to produce many benefits over quasi-likelihood- and likelihood-based frequentist methods. A key benefit of modeling using the Bayesian approach is that it automatically takes into consideration the uncertainty corresponding to model parameter estimation when predicting the parameter estimates. Furthermore, Bayesian methods provide great flexibility in allowing substantive information and new data to be incorporated through an informative prior distribution, otherwise, vague or noninformative prior is assigned [7].

Chen and Dey [10] proposed a Bayesian multivariate logistic model with the use of a scaled multivariate t proposal distribution. Holmes and Leonhard [11] discussed auxiliary variable methods for inference in Bayesian binary and multinomial regression that improves the performance in probit and logistic regression simulation by jointly updating the regression coefficient and the auxiliary variables. In contrast to Reference [10], their approach was exact and totally automatic multivariable sampling schemes for Bayesian binary and polytomous regression methods with full extensions to multinomial regression. Another Bayesian method for joint modeling binary and continuous subunit-specific outcomes was proposed by Dunson [12] with an application to developmental toxicity data. Different methods were proposed for probit and logit models such as the data-augmentation algorithm for missing data, see for example References [13–15]. A similar approach was proposed by Edwards et al. [16] where they implanted the Markov chain Monte Carlo methods (MCMC) technique in which standard identifiability restrictions were excluded. Talhouk et al. [17] proposed Bayesian inference for multivariate probit models with sparse inverse correlation matrices, which takes into consideration the correlation structure between binary observations. More recently, Fasano et al. [18] developed a new variational approximation for posterior probabilities in multivariate probit regression with Gaussian priors. Cao et al. [19] also developed a novel scalable computation of predictive probabilities in probit models with Gaussian process priors.

Aljarallah et al. [20] proposed a classical logistic regression model for analyzing and prediction of Down syndrome and Mental retardation diseases from the Kuwait Medical Genetic Center (KMGC) dataset. This paper is motivated by the need to develop Bayesian methods to model KMGC data. More specifically, it reports on the findings from applying a series of Bayesian univariate prediction algorithms, namely logistic, probit and complementary log-log regression models. Our primary purpose in this paper is to demonstrate how such prediction models can be implemented relatively quickly and easily from a Bayesian perspective using Gibbs sampling MCMC methods in WinBUGS [21]. The importance of MCMC methods lies in their flexibility in specifying complex non-standard models and their ability to easily compute model complexity and fit statistics for non-nested models [22]. Model prediction was evaluated by computing R^2, adjusted R^2, root mean square error (RMSE) and the Bayesian Deviance Information Criterion (DIC).

As the aforementioned methods readily extend to bivariate cases, our secondary purpose in this paper is to model the data by taking the dependence between observations as ignoring the correlations between repeated observations can lead to invalid inferences. A discussion of the various approaches to model correlated binary observations can be found in References [23–25]. However, the models used in all previous analyses have been frequentists. In this paper, bivariate models are implemented from a Bayesian perspective.

2. Materials and Methods

2.1. Study and Dataset Used

The dataset comes from an observational study that was conducted at the Kuwait Medical Genetic Center (KMGC) at the Ministry of Health in Kuwait [20]. The study was considered as the first comprehensive population-based registry whose purpose was to document and assess an extensive spectrum of genetic diseases. The patients were from various nationalities and ethnicities who lived in Kuwait during the period 1979–2009. Over the 30-year period, the total number of patients registered in KMGC was 26,050 with different genetic diseases. However, if patients had partial data, then these patients were excluded from the analysis. For the analysis in this paper, we only consider a complete-case analysis of 17,600 patients, restricted to cases without missing data.

More than 692 disorders were identified which reflect the common and the recorded genetic disorders in Kuwait. Six genetic diseases out of 692 disorders were defined as majors [20]. These disorders were: Down syndrome (DS) (12.1%), Recurrent Pregnancy Loss (RPL) (11%), Multiple Congenital Anomalies (MCL) (10.2%), Mental retardation (MR) (8.4%), Slow Learning (SL) (7.8%) and Cerebral Palsy (CP) (5.9%). For illustration, we only consider the DS and MR datasets here [20].

The KMGC database also provides socio-demographic data (ethnicity, age, gender, nationality, governorate, maternal and paternal age at childbirth and birth order), consanguinity (double first cousin, first cousin, first cousin once removed, etc.), reproductive data (preconception, history of first and second trimester, gravida, reproductive wastage, abortion, still births) and genetic aspects (karyotype, type of chromosomal, aberration, number of affected siblings). Full details of the KMGC study design and methodology are published elsewhere [20].

2.2. Model Development and Validation

Three different models using Bayesian methods were fitted to the data, namely logistic, probit and complementary log-log regression models. All models shown are executed using Bayesian Gibbs sampling MCMC methods in WinBUGS [21], and the relevant WinBUGS code is provided in the Supplementary Material (Code S1).

In each model, the dummy variable which takes on value 1 if the patient has the disease and value 0 otherwise is treated as our dependent variable. Further, both DS and MR diseases are assumed to be associated with a number of covariates at the individual level of response. In this application, only covariates with the greatest discriminatory ability were included in the models. Thus, the covariates considered are listed below:

- Amniotic Fluid—categorical (1 = Yes, 0 = No)
- Complications During Pregnancy—categorical (1 = Yes, 0 = No)
- Ethnicity—categorical (1 = family, 0 = tribe)
- Gestational Age—continuous
- Maternal Age at Childs Birth—continuous
- Nationality—categorical (1 = Kuwaiti, 0 = Non-Kuwaiti)
- Parental Couple Consanguinity—categorical (1 = Yes, 0 = No)
- Pre-conceptional History—categorical (1 = Yes, 0 = No)
- Sex—categorical (1 = female, 0 = male)
- Age—continuous

2.2.1. Model Development

Let

$$Y_i = \begin{cases} 1, & \text{if patient } i \text{ has the disease} \\ 0, & \text{otherwise} \end{cases}$$

where $i = 1, \ldots, n$, and n is the total number of observations. Since Y_i is a binary variable, it has a Bernoulli distribution with parameter $p_i = P(Y_i = 1)$, that is, p_i is the probability of having the disease. Thus,

$$Y_i \sim Bernoulli\ (p_i)$$

Model 1: The logistic regression model is most widely used to predict the patients with the disease. That is,

$$\text{logit}[P(Y_i = 1)|X] = \log\left[\frac{p_i}{1-p_i}\right] = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}, \quad (1)$$

where the X_i's represent individuals' values for the covariates and β's are the k unknown regression parameters in the logistic model.

Model 2: The probit regression model can also be used to predict those patients with the disease. That is,

$$\text{probit}[P(Y_i = 1)|X] = \Phi^{-1}(p_i) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}, \tag{2}$$

Model 3: The third model fitted was a complementary log-log or cloglog model. That is,

$$\text{cloglog}[P(Y_i = 1)|X] = \log(-\log(1 - p_i)) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}. \tag{3}$$

The Bayesian model is completed by assigning prior distributions for all unknown parameters $\beta_0, \beta_2, \ldots, \beta_k$. In the case when no particular prior information is available, non-informative prior distributions are assigned to these parameters. Typically, multivariate normal prior distributions are chosen for these parameters with zero mean and large variance. More specifically, prior distributions were specified as follows:

$$\beta_0, \beta_2, \ldots, \beta_k \sim N\left(0,\ 10^6\right)$$

More details on the choice of the noninformative prior distribution are given in Natarajan and Kass [26].

2.2.2. Model Complexity and Fit

The performance of all models was compared by computing the Bayesian Deviance Information Criterion (DIC) [22]. The DIC is specified by

$$DIC = \overline{D} + P_D$$

where $\overline{D}$ represents the posterior mean deviance and P_D is the effective number of parameters which represents model complexity. The DIC is analogous to Akaike Information Criterion (AIC) [27] and, in the Bayesian framework, it is used to assess model fit penalized for increased model complexity. Spiegelhalter et al. [22] suggest that the best fitting model is defined by the minimum DIC estimates. This criterion has been adopted here for the assessment of model prediction and fit.

2.2.3. Model Prediction

To check the predictive ability of the assumed models, models 1–3 were applied to derive. The predicted probability under the different models is given by the following equations:

Model 1:

$$\hat{p}_i = \frac{\exp(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})}{1 + \exp(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})} \tag{4}$$

Model 2:

$$\hat{p}_i = \Phi(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}). \tag{5}$$

Model 3:

$$\hat{p}_i = 1 - \exp[-\exp(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})]. \tag{6}$$

The predicted probabilities are computed by plugging in the estimated β parameters from each MCMC simulation iteration to the observed values.

The performance of all models was compared by calculating the unadjusted R^2 statistic [28], that is

$$R^2 = 1 - \frac{\log \hat{L}(M_{Full})}{\log \hat{L}(M_{Intercept})} \tag{7}$$

where M_{Full} is the model that contains predictors, whilst $M_{Intercept}$ is the model that contains the intercept only and $\hat{L}$ is the estimated likelihood. McFadden's Formula (7) is similar to the unadjusted R^2 in the ordinary least-squares (OLS) approach in a sense that the log likelihood of $M_{Intercept}$ and M_{Full} are used as total sum of squares and the sum of squared errors, respectively. The ratio of the likelihoods explains the degree of improvement of M_{Full} over $M_{Intercept}$. A likelihood lies within 0 and 1, so log-likelihood is negative. Thus, a small ratio of log-likelihoods means that M_{Full} produces far better fit than $M_{Intercept}$. Upon comparing two models on the same data, as is the case here, McFadden's R^2 (7) would be bigger for the model with the higher likelihood.

Further, we calculated the adjusted R^2 that adjusts for the number of predictors in the model. Specifically,

$$\text{Adjusted } R^2 = 1 - \{\left(1 - R^2\right) * \left[\frac{n-1}{n-k-1}\right]\} \tag{8}$$

where n represents the sample size and k is the total number of covariates in the model. As is the case with the unadjusted R^2, McFadden's adjusted is similar to the adjusted R^2 in the OLS approach by penalizing a model for including lots of covariates.

To test the performance of the three proposed models (outlined above), we compared them based on the calculation of R^2, adjusted R^2, root mean square error criterion and the Bayesian DIC. It is to be noted that Equations (4)–(6) are needed to compute such criterions used. The WinBUGS code (available in the Supplementary Material (Code S1)) provides more details on this.

2.2.4. Bivariate Logistic Regression

We now develop the bivariate logistic regression algorithm for modeling bivariate binomial responses. Bivariate logistic regression has the potential of modeling the marginal probability distribution of the bivariate binary outcomes in addition to modeling the odds ratio that describes the pairwise correlation between the two binary outcomes with respect to some explanatory variables.

Define $\boldsymbol{Y} = (Y_1, Y_2)^T$, where Y_1 and Y_2 take only the values 0 and 1. As usual, we denote "failure" by 0 and "success" by 1. Now define $p_{rs} = P(Y_1 = r, Y_2 = s)$, $r, s = 0, 1$, the joint probabilities, and $p_j = P\left(Y_j = 1\right)$, $j = 1, 2$, the marginal probabilities. It is to be noted that the observations within (different) pairs are correlated (independent). Further information on bivariate binomial data is available elsewhere [29].

The bivariate logistic model described in References [29,30] and later in Reference [22] is defined by modeling the marginal distribution of each of Y_j as well as the odds ratio. The latter is defined as $\psi = p_{00}p_{11}/(p_{01}p_{10})$, and is used to provide the association between the two outcomes. Thus, the model is given by,

$$\begin{aligned} Y_j &\sim Bernoulli(p_j) \\ \text{logit } p_j &= \beta_j^T \boldsymbol{X}, \quad j = 1, 2 \\ \log \psi(x) &= \beta_3^T \boldsymbol{X} \end{aligned}$$

where the $\boldsymbol{X}$ indicates respondents' values for the covariates and β's are the unknown regression coefficients in the bivariate logistic model to be estimated. The joint probability p_{11} can be obtained in terms of p_1, p_2 and ψ, as

$$p_{11} = \begin{cases} \frac{1}{2}(\psi - 1)^{-1}\left\{a \mp \sqrt{a^2 + b}\right\}, & \psi \neq 1 \\ p_1 p_2, & \psi = 1 \end{cases}$$

where a = 1 + $(p_1$ + $p_2)(\psi - 1)$ and b = $-4\psi(\psi - 1)$ $p_1 p_2$. The remaining three probabilities p_{rs} can be straightforwardly obtained from the marginals and p_{11}.

The predicted probability for each observation under the bivariate logistic model is given by

$$\hat{p}_j = \frac{1}{1+\exp\left(\beta_j^T \boldsymbol{X}\right)}, \ j = 1,2$$
$$\psi = \exp(\beta_3^T \boldsymbol{X})$$

As in the univariate case, the predicted probabilities are computed by plugging in the estimated β parameters from each MCMC simulation iteration to the observed values. The relevant WinBUGS code is provided in the Supplementary Material (Code S2).

3. Results

3.1. Study Cohort

The entire KMGC study population consisted of 26,050 patients, 8450 of which had missing observations for one or more of the socio-demographic characteristics, consanguinity, reproductive values and for the DS and/or MR outcome data. Therefore, our complete case analysis was conducted on a group of 17,600 patients (32.4% had missing observations). On average, maternal age at child's birth was 29.18, and children were 14.93 years old. The study sample consisted of a slightly higher proportion of males versus females (55.6% males and 44.4% females). Further, the majority of the patients were Kuwaitis (71.4%) and came from family (77.8%), whilst only 12.7% reported were having complications during pregnancy and 17% reported having pre-conceptional history. The baseline characteristics of the entire population and the complete case cohort, along with full details of the KMGC study methodology, are presented in Reference [20].

3.2. Model Estimation

The implementation of all models using Bayesian methods is shown through its application to the KMGC data mentioned in Section 2. After the burn-in iterations, the Gibbs sampler was scanned for 10,000 iterations to reach convergence. The convergence of the Gibbs sampler was examined using the Gelman and Rubin [31] convergence statistic for two parallel chains with different starting values. The ratio of the within-chain to between-chain variance was then computed. A ratio of about 1 means that convergence had been reached. For the purpose of parameter estimation, a further 25,000 iterations were then followed.

3.3. Model Diagnostics

The results for each model are displayed in Table 1. For each parameter, the posterior mean and associated 95% credible interval are presented. The first three columns show the parameter estimates for the three models (95% credible interval in parentheses) for the DS data. Results revealed that 8 out of the 10 coefficients had credible intervals excluding zero in all three models. Based on the DIC, both the probit and cloglog regression models produced a comparable fit to the data with values of 7448.61 and 7451.21, respectively. As far as the logistic regression model is concerned, the DIC was 7435.98. Overall, the logistic regression model was found to produce the best fit to the data when compared to both probit and cloglog regression models. The last three columns show the parameter estimates for the three models for the MR data. As is the case with DS analysis, eight coefficients had credible intervals excluding zero in all three models. Further, the logistic regression model was found to provide the best fit (DIC = 23,120), when compared to both probit (DIC = 23,180) and cloglog (DIC = 23,190) regression models.

Table 1. Parameter estimates for the three models (95% credible interval in parentheses) for each of the Down syndrome (DS) and Mental retardation (MR) data.

		DS			MR	
Covariate	**Logit**	**Probit**	**Cloglog**	**Logit**	**Probit**	**Cloglog**
Intercept	**−6.868** **(−7.533, −4.68)**	**−0.5829** **(−0.8539, −0.0541)**	**−1.105** **(−1.518, −0.094)**	**−2.282** **(−2.626, −1.859)**	**−0.5294** **(−0.584, −0.428)**	**−0.5971** **(−0.851, −0.296)**
Ethnicity	**−0.140** **(−0.260, −0.021)**	**−0.079** **(−0.143, −0.017)**	**−0.108** **(−0.212, −0.004)**	0.031 (−0.045, 0.105)	0.018 (−0.026, 0.064)	0.019 (−0.035, 0.075)
Sex	0.025 (−0.075, 0.121)	0.015 (−0.036, 0.068)	0.017 (−0.068, 0.102)	**−0.434** **(−0.496, −0.374)**	**−0.267** **(−0.305, −0.229)**	**−0.323** **(−0.370, −0.277)**
Nationality	**−0.637** **(−0.736, −0.538)**	**−0.346** **(−0.399, −0.292)**	**−0.548** **(−0.634, −0.461)**	**0.244** **(0.175, 0.312)**	**0.149** **(0.107, 0.193)**	**0.175** **(0.123, 0.226)**
Parental Couple Consanguinity	**−0.485** **(−0.589, −0.386)**	**−0.257** **(−0.311, −0.204)**	**−0.430** **(−0.520, −0.341)**	**−0.302** **(−0.365, −0.241)**	**−0.188** **(−0.227, −0.149)**	**−0.220** **(−0.266, −0.174)**
Maternal Age at Child's Birth	**0.131** **(0.123, 0.138)**	**0.069** **(0.065, 0.072)**	**0.112** **(0.106, 0.117)**	**0.044** **(0.039, 0.049)**	**0.026** **(0.023, 0.029)**	**0.031** **(0.027, 0.034)**
Pre-conceptional History	−0.069 (−0.212, 0.070)	−0.033 (−0.108, 0.040)	−0.057 (−0.181, 0.065)	**−0.681** **(−0.774, −0.591)**	**−0.419** **(−0.475, −0.364)**	**−0.520** **(−0.593, −0.449)**
Gestational Age	**0.064** **(0.041, 0.088)**	**0.031** **(0.022, 0.043)**	**0.059** **(0.045, 0.074)**	**−4.423** **(−6.908, −3.817)**	**−0.188** **(−0.227, −0.149)**	**0.010** **(0.002, 0.018)**
Amniotic Fluid	**−0.273** **(−0.384, −0.164)**	**−0.146** **(−0.206, −0.085)**	**−0.246** **(−0.345, −0.147)**	0.071 (−0.003, 0.145)	0.044 (−0.0005, 0.087)	0.046 (−0.009, 0.101)
Complications During Pregnancy	**−0.598** **(−0.770, −0.437)**	**−0.316** **(−0.405, −0.228)**	**−0.543** **(−0.694, −0.390)**	**−0.247** **(−0.350, −0.149)**	**−0.152** **(−0.213, −0.091)**	**−0.189** **(−0.267, −0.113)**
Age	**0.006** **(0.001, 0.011)**	**0.003** **(0.001, 0.006)**	**0.002** **(0.001, 0.007)**	**0.013** **(0.010, 0.016)**	**0.008** **(0.006, 0.010)**	**0.008** **(0.006, 0.010)**
R^2	0.1145	0.0778	0.0755	0.0626	0.0594	0.0576
Adjusted R^2	0.114	0.0772	0.0749	0.0621	0.0589	0.0571
RMSE	0.3074	0.3137	0.3141	0.4676	0.4828	0.4833
Overall DIC						
$\overline{D}$	7400.77	7414.02	7416.71	23,100	23,150	23,170
P_D	17.603	17.293	17.246	9.123	11.21	7.344
DIC	7435.98	7448.61	7451.21	23,120	23,180	23,190

Note: DS: Down syndrome; MR: Mental retardation; RMSE: root mean square error; DIC: Deviance Information Criterion. Values reported as posterior means with their 95% credible intervals. Estimates shown in bold are those that have credible intervals excluding zero.

3.4. Model Prediction

To assess the predictive ability of the proposed models, R^2, adjusted R^2 and RMSE were computed for the three models and were also shown in Table 1. We see that for predicting probabilities of observing the disease for the cohort of individuals with DS and/or MR, the logistic regression model performed best under all criterions used, with R^2 (0.1145), adjusted R^2 (0.114) and RMSE (0.3074) for DS, and with R^2 (0.0626), adjusted R^2 (0.0621) and RMSE (0.4676) for MR.

3.5. Bivariate Case

It is more likely that both DS and MR diseases are associated with a number of covariates at the respondent level. Therefore, a bivariate analysis is conducted with each of the pre-defined covariates in order to explore the significant related factors as well as to study the association between the two diseases. Table 2 describes the results of the bivariate logistic regression analysis, reflecting the association between DS and MR. Results revealed that 7 and 8 out of the 10 selected covariates were significantly associated with DS and MR respectively, whilst none of the selected covariates was associated with the interaction between DS and MR.

Table 2. Bayesian odds ratio estimates (95% credible interval in parentheses) for the bivariate logistic regression model for both DS and MR diseases along with their association.

	Bayesian Estimates		
	Odds Ratio (95% Credible Interval)		
Covariate	**DS**	**MR**	**Association**
Constant	**0.0005** **(0.0002, 0.0010)**	**0.1040** **(0.0634, 0.1606)**	0.9196
Amniotic Fluid	1.0960 (0.8037, 1.4780)	**1.7800** **(1.4580, 2.1660)**	1.6641
Complications During Pregnancy	**0.5916** **(0.4985, 0.6948)**	**0.7668** **(0.6938, 0.8452)**	0.9273
Ethnicity	**0.8452** **(0.7508, 0.9496)**	1.0350 (0.9585, 1.1150)	1.0019
Gestational Age	**1.0600** **(1.0390,1.0830)**	1.0100 (0.9991, 1.0210)	1.0129
Maternal Age at Child’s Birth	**1.1400** **(1.1320, 1.1490)**	**1.0450** **(1.0400, 1.0500)**	1.1482
Nationality	**0.5261** **(0.4754, 0.5806)**	**1.2740** **(1.1880, 1.3640)**	0.7584
Parental Couple Consanguinity	**0.6206** **(0.5605, 0.6854)**	**0.7430** **(0.6970, 0.7910)**	1.0681
Pre-conceptional History	0.9662 (0.8379, 1.1070)	**0.5044** **(0.4597, 0.5518)**	0.9208
Sex	1.0290 (0.9328, 1.1320)	**0.6499** **(0.6102, 0.6912)**	0.9605
Age	**1.0070** **(1.0020, 1.0120)**	**1.0120** **(1.0090, 1.0160)**	1.0747

Note: DS: Down syndrome; MR: Mental retardation. Values reported as odds ratios with their 95% credible intervals. Estimates shown in bold are those that have credible intervals excluding one.

The odds ratios (OR) with the corresponding 95% posterior credible intervals for the intercept and each of the covariates are presented in Table 2. Although Amniotic Fluid was not significantly associated with DS, it had some significant impact on MR in the sense that subjects who had Amniotic Fluid were 1.78 times more likely to have MR. As far as

Complications During Pregnancy is concerned, we see that it had a significant impact on both DS and MR. That is, subjects with Complications During Pregnancy were less likely to get both DS and/or MR. This is also the case with Parental Couple Consanguinity. Ethnicity was not significantly associated with MR but had some significant impact on DS in the sense that subjects from family were less likely (OR = 0.8452) to have DS. Regarding Maternal Age at Child's Birth, results revealed that it also had a significant impact on both DS and MR. That is, subjects with higher Maternal Age at Child's Birth were more likely to get both DS (OR = 1.1400) and/or MR (OR = 1.0450). The situation with Nationality is challenging somehow. It can be seen that Nationality had a significant impact on both DS and MR. However, Kuwaitis were less likely (OR = 0.5261) to get DS but 1.274 times more likely to have MR. Further, gender was not significantly associated with DS, but it had some significant impact on MR. More specifically, females were less likely (OR = 0.6499) to have MR. To this end, it is apparent from Table 2 that age had a significant impact on both DS and MR, where it is evident from the results that older subjects were more likely to get both DS (OR = 1.0070) and/or MR (OR = 1.0120).

4. Discussion

In this paper, we have developed a series of Bayesian models for DS and MR of the KMGC dataset. Models were used in two distinct ways; firstly, using logistic, probit and complementary log-log regression models, and, secondly, using bivariate logistic regression to study the association between the two diseases in question. In the univariate analysis, the logistic regression model was found to provide the best fit to both DS and MR data under all criterions used when compared to both probit and cloglog regression models. The logistic regression model performed best, with R^2(0.1145), adjusted R^2(0.114), RMSE (0.3074) and DIC (7435.98) for the DS data, and R^2(0.0626), adjusted R^2(0.0621), RMSE (0.4676) and DIC (23,120) for the MR data (see Table 1). In the bivariate case, it is evident from the results that 7 and 8 out of the 10 selected covariates were significantly associated with DS and MR respectively, whilst none of the selected covariates were associated with the interaction between DS and MR (see Table 2).

All models presented here were executed using Bayesian Gibbs sampling MCMC methods in WinBUGS. These methods have key advantages over the classical approach given that they permit great flexibility in (1) handling complex non-standard models, (2) allowing model fit and complexity to be assessed straightforwardly and (3) automatically incorporating all parameter estimation uncertainty into the results. Additionally, the DIC that is used for model selection is straightforward to calculate in a MCMC analysis. Bayes factors would have been a great alternative criterion for model selection, though, in order to assess the relative ability of the different models in predicting the data. However, in comparison to the DIC, implementing Bayes factors needs the specification of informative prior distributions, which we did not use here, and is the subject of further work.

Limitations of our study include the use of vague prior distributions for all the presented models. In a Bayesian approach, it is straightforward to incorporate substantive external information and data via the prior distribution, for example, the use of conjugate informative priors when highly strong interactions matter among covariates. An additional limitation is identifying the best prior for the parameters in the bivariate model. In the analysis presented here, a large variance normal distribution is defined in order to overcome this problem. Although this was intended to be noninformative, it would be of great interest to perform sensitivity analysis by assigning a range of prior distributions. A further limitation is that the available dataset contains most of the important covariates; however, the effect of reproductive data (history of first and second trimester, gravida, reproductive wastage, abortion, still births) and genetic aspects (karyotype, type of chromosomal, aberration, number of affected siblings) on DS and/or MR cannot be overlooked. Data pertaining to other disorders such as Recurrent Pregnancy Loss, Multiple Congenital Anomalies, Slow Learning and Cerebral Palsy can also be analyzed using the developed bivariate binary logistic regression model. All of the above are the subjects of further work.

An additional concern is around individuals with missing covariate data. Note that a complete case analysis was only considered here. The overall percentage of missing values was 32.4%. We did not notice any significant differences in either socio-demographic characteristics or the chances of getting DS and/or MR between the overall cohort and the complete case dataset. Missing observations are usually ignored in many standard regression software packages. However, within WinBUGS, missing observations are treated as unknown quantities to be estimated by the model. This implies that missing data from respondents, which is ignored in the current analysis, could certainly be included, and hence more data would be available to fit and validate the model [32,33].

This paper has proposed a series of Bayesian models for modeling and predicting DS and MR diseases of the KMGC dataset. The Bayesian analyses presented here have illustrated how such prediction models can be implemented relatively quickly and easily from a Bayesian perspective using Gibbs sampling MCMC methods in WinBUGS. Such models may provide significant information (such as estimation of future chance of getting DS and/or MR as well as the association between the two diseases) for healthcare decision-makers in the public organizations such as the Ministry of Health in Kuwait and private sector firms on the planning of future services and budgets.

Supplementary Materials: The following are available online at https://www.mdpi.com/2227-7390/9/3/248/s1, Code S1: WinBUGS code for Logistic Regression Model (Model 1), Code S2: WinBUGS code for Bivariate Logistic Regression Model.

Author Contributions: Conceptualization, R.A. and S.A.K.; methodology, R.A. and S.A.K.; software, S.A.K.; validation, R.A. and S.A.K.; formal analysis, S.A.K.; investigation, R.A. and S.A.K.; resources, R.A. and S.A.K.; data curation, R.A.; writing—original draft preparation, R.A. and S.A.K.; writing—review and editing, R.A. and S.A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the leading author. The data are not publicly available due to privacy.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zeger, S.L.; Liang, K.Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **1986**, *42*, 121–130. [CrossRef]
2. Yi, G.Y.; Cook, R.J. Marginal methods for incomplete longitudinal data arising in clusters. *J. Am. Stat. Assoc.* **2002**, *97*, 1071–1080. [CrossRef]
3. Fang, D.; Sun, R.; Wilson, J.R. Joint modeling of correlated binary outcomes: The case of contraceptive use and HIV knowledge in Bangladesh. *PLoS ONE* **2018**, *13*, e0190917. [CrossRef]
4. El-Sayed, A. Modeling Multivariate Correlated Binary Data. *Am. J. Theor. Appl. Stat.* **2016**, *5*, 225–233. [CrossRef]
5. Lee, K.; Joo, Y.; Yoo, J.K.; Lee, J.B. Marginalized random effects models for multivariate longitudinal binary data. *Stat. Med.* **2009**, *28*, 1284–1300. [CrossRef]
6. Stiratelli, R.; Laird, N.; Ware, J.H. Random-effects models for serial observations with binary response. *Biometrics* **1984**, *40*, 961–971. [CrossRef]
7. O'Brien, M.; Dunson, B. Bayesian multivariate logistic regression. *Biometrics* **2004**, *60*, 739–746. [CrossRef]
8. Fitzmaurice, G.; Laird, N. A likelihood-based method for analysing longitudinal binary responses. *Biometrika* **1993**, *80*, 141–151. [CrossRef]
9. Qaqish, B.; Ivanova, A. Multivariate logistic models. *Biometrika* **2006**, *93*, 1011–1017. [CrossRef]
10. Chen, M.H.; Dey, D.K. Bayesian modeling of correlated binary responses via scale mixture of multivariate normal link functions. *Sankhya* **1998**, *60*, 322–343.
11. Holmes, C.; Leonhard, H. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Anal.* **2006**, *1*, 145–168. [CrossRef]
12. Dunson, D.B.; Chen, Z.; Harry, J.A. Bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics* **2003**, *59*, 521–530. [CrossRef]

13. Frühwirth-Schnatter, S.; Frühwirth, R. Data augmentation and MCMC for binary and multinomial logit models. In *Statistical Modelling and Regression Structures*; Kneib, T., Tutz, G., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 111–132.
14. Chopin, N.; Ridgway, J. Leave Pima Indians alone: Binary regression as a benchmark for Bayesian computation. *Stat. Sci.* **2017**, *32*, 64–87. [CrossRef]
15. Polson, N.G.; Scott, J.G.; Windle, J. Bayesian inference for logistic models using pólya–gamma latent variables. *J. Am. Stat. Assoc.* **2013**, *108*, 1339–1349. [CrossRef]
16. Edwards, Y.D.; Allenby, G.M. Multivariate analysis of multiple response data. *J. Mark. Res.* **2003**, *40*, 321–334. [CrossRef]
17. Talhouk, A.; Doucet, A.; Murphy, K. Efficient bayesian inference for multivariate probit models with sparse inverse correlation matrices. *J. Comput. Graph. Stat.* **2012**, *3*, 739–757. [CrossRef]
18. Fasano, A.; Durante, D.; Zanella, G. Scalable and accurate variational Bayes for high-dimensional binary regression models. *arXiv* **2019**, arXiv:1911.06743.
19. Cao, J.; Durante, D.; Genton, M. Scalable computation of predictive probabilities in probit models with Gaussian process priors. *arXiv* **2020**, arXiv:2009.01471v2.
20. Al-Jarallah, R.; Al-Awadi, S.; Bastaki, L.; Marafi, M. *Genetic Diseases in State of Kuwait: A Statistical Approach*; Technical Report; University of Kuwait: Kuwait City, Kuwait, 2012.
21. Spiegelhalter, D.; Thomas, A.; Best, N.; Lunn, D. *WinBUGS User Manual: Version 1.4*; MRC Biostatistics Unit: Cambridge, UK, 2003.
22. Spiegelhalter, D.J.; Best, N.G.; Carlin, B.P.; van der Linde, A. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B* **2002**, *64*, 583–616. [CrossRef]
23. Le Cessie, S.; Van Houwelingen, J.C. Logistic regression for correlated binary data. *Appl. Stat.* **1994**, *43*, 95–108. [CrossRef]
24. Sheu, C.-F. Regression analysis of correlated binary outcomes. *Behav. Res. Methods Instrum. Comput.* **2000**, *32*, 269–273. [CrossRef] [PubMed]
25. Bhatnagar, S.R.; Atherton, J.; Benedetti, A. Comparing alternating logistic regressions to other approaches to modelling correlated binary data. *J. Stat. Comput. Simul.* **2015**, *85*, 2059–2071. [CrossRef]
26. Natarajan, R.; Kass, R.E. Reference Bayesian methods for generalized linear mixed models. *J. Am. Stat. Assoc.* **2000**, *95*, 227–237. [CrossRef]
27. Akaike, H. Information theory and an extension of the maximum likelihood principle. In Proceedings of the 2nd International Symposium Information Theory, Tsahkadsor, Armenia, 2–8 September 1971; Petrov, B.N., Caski, F., Eds.; Akademiai Kiado: Budapest, Hungary, 1973; pp. 267–281.
28. McFadden, D. Conditional logit Analysis of Qualitative choice behavior. In *Frontiers of Econometrics*; Zarembka, P., Ed.; Academic Press: New York, NY, USA, 1974; pp. 105–142.
29. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*, 2nd ed.; Chapman & Hall: London, UK, 1989.
30. Palmgren, J. *Regression Models for Bivariate Binary Responses*; Technical Report No. 101; Department of Biostatistics, University of Washington: Seattle, WA, USA, 1989.
31. Gelman, A.; Rubin, D. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **1992**, *7*, 457–472. [CrossRef]
32. Sweeting, T.J.; Kharroubi, S.A. Application of a predictive distribution formula to Bayesian computation for incomplete data models. *Stat. Comput.* **2005**, *15*, 167–178. [CrossRef]
33. Kharroubi, S.A.; Sweeting, T.J. Posterior simulation via signed root log-likelihood ratios. *Bayesian Anal.* **2010**, *5*, 787–816. [CrossRef]