Computation and application of likelihood prediction with generalized linear and mixed models.

Md Moudud Alam*

October 13, 2010.

Abstract

This paper presents the computation of likelihood prediction with the generalized linear and mixed models. The method of likelihood prediction is briefly discussed and approximate formulae are provided to make easy computation of the likelihood prediction with generalized linear models. For complicated prediction problems, simulation methods are suggested. An R add-in package is accompanied to carry out the computation of the predictive inference with the generalized linear and mixed models. The likelihood prediction is applied to the prediction of the credit defaults using a real data set. Results show that the predictive likelihood can be a useful tool to predict portfolio credit risk.

Key words: Predictive likelihood, Profile predictive likelihood, Coverage interval, Future value prediction, Credit risk prediction, R-package.

1 Introduction

In real world application we often want to make prediction. For example, while processing a credit application the credit issuer needs to predict the probability of the borrower being able to repay the loan. However, the computational procedure to produce a predictive inference is rarely available in the existing software. To illustrate the computational issue let us consider the statistical software (R Development Core Team 2009) as an example. R has a generic function, predict(), which is supposed to produce predictive inferences. When it is applied to some object class, e.g. lm, one gets predictive inferential statistics computed. However, the ‘predict’ method is not implemented for all statistical model objects in R. And when it is implemented, it might produce merely a fitted value not a predicted value. Note that fitted value refers to the expected outcome within the observations actually used for the model fitting while the predicted values refer to those for the future observations.

In R, the glm function implements generalized linear models (McCullagh & Nelder 1989) while lm (package: nlme; Pinheiro & Bates (2000)), lmer (package: lme4; Bates

---

*School of Technology and Business Studies, Dalarna University and Swedish Business School, Örebro University. Contact: Dalarna University, SE 781 88 Borlänge, Sweden; maa@du.se.
(2005)) and hglm (package: hglm; Rönnegård, Shen & Alam (2010), 2010) implement linear and generalized linear mixed models (GLMM; McCulloch and Searle, 2001). Though, there are predict methods in R for glm and lme, they produce the fitted values, not the predicted values. The predicted values are sometimes but not necessarily the same as the fitted values and the uncertainty is obviously different (Bjørnstad 1996, Patel 1989). Furthermore, the predict method is not available for lmer and hglm object class (see appendix A.3 for a complete list of predict methods). The unavailability of predictive inference is not particular to R, the lack is similar if one choose an alternative software such as SAS (Littell, Milliken, Stroup & Wolfinger 1996).

The unavailability of computational procedures for prediction may not be attributed to their limited usefulness, rather it is due to the lack of theoretical research and empirical evidences on the precision and the applicability of the predictive inferential framework for the generalized linear and mixed models. In this paper we present computational details of the likelihood prediction for the GLM and GLMM. We propose three different ways to approximate the computation of the likelihood prediction for GLM and GLMM and compare their performances by Monte-Carlo simulations. We also give an accompanying R add-in package to carry out the predictive inference.

The procedures are illustrated by a standard text book example (Davison & Hinkley 1997, PP. 342–345) on predicting aids reports in England and Wales and an application of the predictive inference using a real data set on credit defaults. Assuming that the credit defaults can be modelled by using a generalize linear mixed models, we compare the performance of the different strategies to predict future credit risk. By simulation and by application to a real data set on credit defaults we show that the predictive likelihood (Bjørnstad 1990) is a useful tool to predict credit defaults and to compute the uncertainty. It has been shown elsewhere that both the uncertainty in the model parameters and the uncertainty in the future independent variables can be handled in the framework of predictive likelihood. That enables us to make a valid statement of uncertainty about the credit defaults prediction.

The rest of the paper is organized as follows. Section two gives a short introduction to predictive likelihood. Section three describes how the predictive likelihood can be applied to generalized linear models and demonstrates its superiority over the naive frequentist prediction by means of simulation. Section four illustrates the use of pred package to carry out predictive computation for GLM. Section five shows the usefulness of likelihood prediction in predicting credit risk using real data sets from two Swedish banks. Section six concludes.

2 Likelihood prediction

Let us denote the observed response with $y$, future (not observed) response as $y^*$, observed covariate $X$ and unobserved covariate $X^*$ and the vector of model parameters as $\theta = (\theta_1, \theta_2, ..., \theta_p)$. Note that we use superscript asteric (e.g. $X^*$) to indicate unobserved part of an observable quantity (see Alam (2010b) for a detailed description of the notation system). Then the likelihood principle (Berger & Wolpert 1988, Bjørnstad 1990)) states
that all the sample information about \( \theta \), \( X^* \) and \( y^* \) is contained in the joint likelihood

\[
L(y^*, \theta, X^* | y, X) = f(y, y^*, X^* | X, \theta)
\]

Unlike the estimation problem, the inferential interest of the prediction problem centered on \( y^* \) while all other unobservable, both parameters and variables, are nuisance parameters and variables. However, the likelihood principle does not state as to how the information about \( y^* \) should be extracted from the joint likelihood (Berger and Wolpert, 1988). A direct maximization of (1) w.r.t. \( y^* \), \( \theta \) and \( X^* \) can be disastrous due to the presence of a huge number nuisance quantities (Berger and Wolpert, 1988). Thus some special technique is required to deal with the nuisance parameters and variables in order to extract optimal information on \( y^* \) from the joint likelihood.

The first thing one can do is to integrate out the nuisance variables and other random quantities from the joint likelihood. This is justified by the random quantities having a distribution and hence we can integrate them out from the joint distribution to get the marginal distribution of \( y | X, \theta \) which is free from the random quantities. It is also noted in the literature that replacing \( \theta \) with its observed data maximum likelihood estimator produces misleadingly precise inference on \( y^* \) (Bjørnstad 1996). A survey on the different ways to produce predictive likelihood inference is presented in (Bjørnstad 1990). Based on an extended definition of the likelihood principle (Bjørnstad 1996), five out of 14 different predictive likelihoods surveyed in (Bjørnstad 1990) are found to be proper predictive likelihood. Alam (2010b) pointed out that all the proper predictive likelihoods are not equivalent in general nor can all of them be implemented in every situations. Comparing the theoretical properties and applicability of alternative predictive likelihoods, Alam (2010b) suggested using adjusted profile likelihood technique. Let, \( \hat{\theta} \) be the maximum likelihood estimator (MLE) of \( \theta \) based on only observations on and associated with \( y \) and \( \hat{\theta}^* \) be the MLE based on joint likelihood (based on both observed and unobserved data). Then the adjusted profile predictive likelihood is given by

\[
L(y^* | X) = L\left(z | \theta = \hat{\theta}^*, X^* = \hat{X}^*, y, X\right) \left| \mathcal{I}\left(\hat{\theta}^*, \hat{X}^*\right) \right|^{-\frac{1}{2}}
\]

where \( \mathcal{I}\left(\hat{\theta}^*, \hat{X}^*\right) \) is the observed Fisher’s information obtained from the joint log-likelihood. Davison (1986) showed that (2) is the approximate Bayesian posterior predictive density with flat prior. A computational implication of the relationship is that we can use existing Markov-Chain Monte-Carlo simulation procedures, e.g. those implemented in Bugs (Spiegenhalter, Thomas, Best & Lunn 2007), to carry out the predictive inference.

### 3 Predictive inference with generalized linear and mixed models

Let us assume that the responses, \( y = \{y_i\} \) \((i = 1, 2, ..., n)\), are independently distributed as exponential family, given the covariates \( X_i = (1, X_{i1}, X_{i2}, ..., X_{ip}) \) that is \( f(y_i | X_i, \theta_i) = \exp \left[ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right] \) where \( \theta \) is the canonical parameter and \( a(\phi) \) is the dispersion
The influence of $X$’s on $y$’s are postulated as

$$E(y_i|X_i) = \mu_i$$

$$g(\mu_i) = \eta_i = X_i\beta$$

The above specifies a generalized linear model. Incorporating an additional $j$ dimension (cluster) and a random effect, $u_j$, term in $\eta$ above that is writing $g(\mu_{ij}|u_j) = \eta_i = X_{ij}\beta + u_j$ where $u_j \sim N(0, \sigma_u^2)$ we obtain a generalized linear mixed model. The inferential goal is to produce a set of predictive statistics on the future responses, $y^* = \{y^*_i\}$ ($i = n + 1, n + 2, \ldots, n + m$), after observing $y_i, X_i$. Denote the observed $X_i$ associate with $y_i$ as $\tilde{X}_i$

Assume that the dispersion parameter to be known, which is true only for the Poisson and the binomial family and consider the case of canonical link, for simplicity. The adjusted profile likelihood for the GLM is given by

$$L(y^*|X_F, y) = \exp \left[ \frac{y_T^* \hat{\theta}^* - 1^T \text{diag} \left\{ b(\hat{\theta}^*_{ij}) \right\}}{a(\phi)} + c(y_F, \phi) \right] \left| \frac{1}{a(\phi)} X_F^T W X_F \right|^{-\frac{1}{2}}$$

(3)

where $y_F = (y^T, y^{**T})^T$, $\hat{\theta}^*_i$ is the MLE of $\theta_i$ based on $y_F$ and $X_F$ is the design matrix consisting of $X$ and $\tilde{X}$ and $W$ is the diagonal weight matrix. A direct maximization of (3) or finding a normalizing constant to get (3) a density (or mass) function is a challenging task. Especially, for $y$ being discrete, (3) has to be evaluated for every possible combination of $y^*$. However, we can apply a simulation based computation to draw random sample of $y^*$ from the un-normalized density $L(y^*|X)$. Its connection to Bayesian posterior density allows us to use Bayesian posterior predictive simulation for $y^*$. In absence of any computational method, it can be implemented through Markov-Chain Monte-Carlo techniques. However, using normal approximation to posterior distribution of the parameters we have a first approximation method which we refer to as approximate predictive simulation.

### 3.1 Approximate predictive simulation

We treat (3) as an un-normalized Bayesian posterior predictive distribution with flat prior. Since the MLE of $\theta$ is also the posterior mode, we estimate the model parameters using formal likelihood method on the observed data only. Let, $\hat{\theta}$ be the MLE and $\mathcal{I}(\hat{\theta})$ is the information matrix for $\theta$. Assume that the posterior distribution of $\theta$ is symmetric around $\hat{\theta}$ and can be approximated by a normal density, the predictive simulation can be carried out in the following way.

1. Draw a random sample on $\hat{\theta}^*$ from $\hat{\theta}^* \sim N\left(\hat{\theta}, \mathcal{I}(\hat{\theta})\right)$

2. Draw a random sample on $y^*$ from $y^* \sim f\left(y^*|\theta = \hat{\theta}^*\right)$

3. Summarize the distribution of the simulated $y^*$ as the predictive distribution
This algorithm approximates the Bayesian posterior as long as the normal approximation to the posterior distribution of $\theta$ holds. It also has a frequentist explanation. If $\theta$ were known, then the predictive distribution of $y^*$ would be $f(y^* | \theta)$. Substituting $\hat{\theta}$ for $\theta$ in $f(y^* | \theta)$ would produce a naive predictor (Alam 2010) which does not take uncertainty in $\hat{\theta}$ into account. Since the $\hat{\theta}$ is estimated from the observed data, after correcting for the uncertainty for $\hat{\theta}$, the predictive distribution should be $p(y^*) = \int f(y^* | \hat{\theta}) p(\hat{\theta}) d\hat{\theta}$ where the $p(\hat{\theta})$ is the current estimate of uncertainty about the true parameter $\theta$. Thus the simulation method is simulating draws from $p(y^*)$ my treating $p(\hat{\theta})$ as a normal density.

Besides the simulation based approach we can also use other approximation method to compute predictive statistics from (3). Using the first order Taylor’s approximation to the score function of joint likelihood we obtain our second approximation which is given next.

3.2 First order approximation

In order to obtain an analytical expression for the predictive likelihood (3) we need to know $\hat{\theta}_i$ where $\hat{\theta}_i = X_i \hat{\beta}$ for a GLM with canonical link. Thus we need to know the analytical expression for $\hat{\beta}^*$ which has a close form expression only for the Gaussian family with identity link. In other cases $\hat{\beta}^*$ is computed through an iterative procedure because analytical expression of $\hat{\beta}^*$ as a function of the observed data is not available. However, the estimate of $\hat{\beta}^*$ is obtained by equating the score function of the joint likelihood, denoted by $l^*$, to zero. Hence we obtain $\hat{\beta}^*$ from

$$\frac{\partial l^*}{\partial \beta} = 0$$

Using a first order Taylor’s approximation to $\frac{\partial l^*}{\partial \beta}$ around $\hat{\beta}$ we obtain

$$\left( \frac{\partial l^*}{\partial \beta} \right)_{\beta = \hat{\beta}} + \left( \frac{\partial^2 l^*}{\partial \beta \partial \beta^T} \right)^T (\beta - \hat{\beta}) \approx 0$$

$$\Rightarrow \hat{\beta}^* \approx \hat{\beta} - \left[ \left( \frac{\partial^2 l^*}{\partial \beta \partial \beta^T} \right)^{-1} \frac{\partial l^*}{\partial \beta} \right]_{\beta = \hat{\beta}}$$  \hspace{1cm} (4)

The exact mathematical expression for $\frac{\partial^2 l^*}{\partial \beta \partial \beta^T}$ and $\frac{\partial l^*}{\partial \beta}$ are well-known for the GLMs (see e.g. Wand, 2002) and are given, for canonical link, by

$$\frac{\partial^2 l^*}{\partial \beta \partial \beta^T} = - \frac{1}{a(\phi)} X_F^T \text{diag} \{ b''(X_F \beta) \} X_F$$

and

$$\frac{\partial l^*}{\partial \beta} = \frac{1}{a(\phi)} X_F^T (y_F - b'(X_F \beta))$$
Substituting in (4) we obtain
\[ \hat{\beta}^* \approx \hat{\beta} + \left[ (X_F^T \text{diag} \{ b''(X_F \hat{\beta}) \} X_F)^{-1} X_F^T (y_F - b'(X_F \hat{\beta})) \right]_{\beta=\hat{\beta}} \] (5)

A similar expression can also be given for the non-canonical links but the second term in the right hand side of (5) becomes more messy in that case. Substituting \( \hat{\beta}^* \) in (3) we obtain an approximate predictive likelihood for \( y^* \). The remaining task is either to maximize \( L(y^* | X, y) \), if sensible, or to normalize it to make it a probability density (or a mass) function.

For Gaussian GLM with identity link we have
\[ \hat{\beta}^* \approx \hat{\beta} + (X_F^T X_F)^{-1} X_F^T (y_F - X_F \hat{\beta}) \]
\[ \Rightarrow y_F - X_F \hat{\beta}^* = y_F - X_F \hat{\beta} - X_F (X_F^T X_F)^{-1} X_F^T (y_F - X_F \hat{\beta}) \]
\[ \Rightarrow y_F - X_F \hat{\beta}^* = (I - X_F (X_F^T X_F)^{-1} X_F^T) (y_F - X_F \hat{\beta}) \]
\[ \Rightarrow (y_F - X_F \hat{\beta})^T (y_F - X_F \hat{\beta}^*) = y_F^T (I - X_F (X_F^T X_F)^{-1} X_F^T) y_F \]

Note that \( (y_F - X_F \hat{\beta})^T (y_F - X_F \hat{\beta}^*) \) determines the predictive likelihood uniquely for the Gaussian GLM with known dispersion parameter. In order to obtain the exact predictive distribution we make use of the following lemma (Lemma 1) due to Eaton & Sunderth (1998).

**Lemma 1:** Denote \( S_F = y_F^T (I - X_F (X_F^T X_F)^{-1} X_F^T) y_F \), \( \hat{\beta} = (X^T X)^{-1} X^T y \) where, \( X \) is the design matrix associated with the observed \( y \), \( \hat{X} \) be the design matrix associated with \( y^* \) and \( S = (y - X \hat{\beta})^T (y - X \hat{\beta}) \) then the following identity holds
\[ S_F = S + (y^* - \hat{X} \hat{\beta})^T (I + \hat{X} (X^T X)^{-1} \hat{X}^T)^{-1} (y^* - \hat{X} \hat{\beta}) \]

With the help of Lemma 1, we get the predictive density of \( y^* \) for the Gaussian GLM with identity link as
\[ L(y^* | X_F, y) \propto \exp \left[ -\frac{1}{2a(\phi)} (y^* - \hat{X} \hat{\beta})^T (I + \hat{X} (X^T X)^{-1} \hat{X}^T)^{-1} (y^* - \hat{X} \hat{\beta}) \right] \] (6)

that is the predictive distribution of \( y^* \) is multivariate normal with mean \( \hat{X} \hat{\beta} \) and variance \( a(\phi) (I + \hat{X} (X^T X)^{-1} \hat{X}^T) \) which is exactly the same as the one given in Bjørnstad (1996). Thus the first order approximation is exact for the Gaussian GLM with identity link. Note that the mean prediction from \( L(y^* | X_F, y) \) is also the BLUP for the above Gaussian case.

The approximation method above and the one in section 3.1 are computationally too extensive since in absence of a general solution for the normalizing constant for \( L(y^* | X_F, y) \) we have to rely on Markov-Chain Monte-Carlo techniques. However, if a point predictor and a measure of uncertainty of the point predictor is concern, we can use the following, third, approximation method.

---

1Substituting \( E \left( \frac{\partial y^*}{\partial \beta^*} \right) \) for \( \frac{\partial^2 y^*}{\partial \beta^*} \) in (4), one can also obtain (5) for the non-canonical links.
3.3 Approximate point prediction

From the GLM theory we know that the maximum likelihood estimation of $\beta$ is equivalent to iteratively solving a weighted least square regression (McCullagh & Nelder 1989). This is computationally equivalent to the ML estimation from a multivariate normal model. Thus, using the normal approximation, the joint likelihood of $\xi_i = \eta_i + (y_i - \mu_i) \frac{\partial \mu_i}{\partial \eta_i}, \quad i = 1, 2, ..., (n + m)$ at the $i^{th}$ iteration of the estimation procedure can be expressed as

$$L^*(\xi | X_F, y_F) = - \frac{1}{(2\pi |W|)^{(n+m)/2}} \exp \left[ - \frac{1}{2} (\xi - X_F\beta)^TW^{-1}(\xi - X_F\beta) \right]$$

where $W$ is the so called diagonal weight matrix such that $w_{ii}^{-1} = V(\mu_i) \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2$ and $V(\mu_i)$ is the variance function. The normal linear model theory gives us the best linear unbiased predictor (which is also the likelihood predictor) as $\hat{\xi} = \hat{X}\hat{\beta}$ where, $\hat{\beta} = \hat{\beta}^*$ is the ML estimator of $\beta$ obtained from the observed data only. But, $E\left((y_i - \mu_i) \frac{\partial \mu_i}{\partial \eta_i}\right) = 0$, in turn, produces the likelihood predictor for $y^*$ as

$$\hat{y}_i^* = \hat{\mu}_i^* = g^{-1}\left(\hat{X}_i\hat{\beta}\right); \quad i = n + 1, ..., n + m$$

This is also a naive predictor. However, we do not use the naive estimate of variance of $\hat{y}_i^*$ which would be $Var\left(\hat{y}_i^*\right) = a\left(\phi\right)V(\hat{\mu}_i^*)$. Instead, the variance of $\hat{y}_i^*$ is calculated in the following way

$$Var\left(\hat{y}^*\right) = E_y\left(Var\left(y^*|\beta\right)\right) + Var_y\left(E\left(y^*|\beta\right)\right)$$

$$\Rightarrow Var\left(\hat{y}^*\right) = E\left(V\left(\hat{\mu}^*\right)a\left(\phi\right)\right) + Var\left(g^{-1}\left(\hat{X}\hat{\beta}\right)\right)$$

Using the delta method to calculate $Var\left(g^{-1}\left(\hat{X}\hat{\beta}\right)\right)$, substituting $V(\hat{\mu}^*)a\left(\hat{\phi}\right)$ for $E\left(V(\hat{\mu}^*)a\left(\phi\right)\right)$ we obtain the sample estimate of the standard error (SE) of $\hat{y}^*$ as

$$\hat{se}\left(\hat{y}_i^*\right) \approx \sqrt{V\left(\hat{\mu}_i^*\right)a\left(\hat{\phi}\right) + a\left(\hat{\phi}\right)\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2_{|\mu_i = \hat{\mu}_i^*}} \hat{X}_i(X^TWX)^{-1}\hat{X}_i^T$$

From (6) we note that (8) and (9) provides the exact mean prediction and predictive standard deviation for the Gaussian model with identity link. For binomial response, (9) simplifies into

$$\hat{se}\left(\hat{y}_i^*\right) = \sqrt{\hat{\mu}_i^* (1 - \hat{\mu}_i^*) + (\hat{\mu}_i^* (1 - \hat{\mu}_i^*))^2 \hat{X}_i(X^TWX)^{-1}\hat{X}_i^T}$$

where the first term in the $\hat{se}\left(\hat{y}_i^*\right)$ is well-recognized as the binomial variance and the second term being the additional uncertainty due to the estimation of $\beta$.

In matter of fact, the estimation of GLMM is also carried out in penalized quasi likelihood (Breslow & Clayton 1993, PQL) or pseudo likelihood (Wolfgang & O’Connell 1993) where an iterative weighted least square regression on the modified response $\xi$ is
used to get an updated estimate for $\beta$. Thus, treating the variance components as known, the same approximation for GLM can be extended to produce the predicted values for $y^*_i$ i.e. $\hat{y}^*_i = g^{-1}(\hat{\eta}_i)$. This works under the assumption that the variance components are known which is often not the case. Therefore, we recommend using simulation based prediction for GLMM.

4 Monte-Carlo study

In order to compare the performance of the naive predictor (NP), Likelihood predictor (LP) and several approximations to LP we conduct a small simulation experiment with a logistic GLM. We generate 60 observations from a binomial $(1, \mu_i)$ distribution where, \[ \log \left( \frac{\mu_i}{1-\mu_i} \right) = \eta_i, \quad i = 1, 2, ..., 60 \] and $\eta_i = X_i \beta$. The first column of the design matrix, $X$, is 1 and the rests are independent simulated draws from $N_5(0, diag \{0.7, 0.5, 1, 0.8, 1.2\})$. The model parameters are arbitrarily chosen as $\beta = (-0.6, 2, 1.1, -0.7, 0.8)$. With the first 40, out of the 60, observations we fit a logistic model and estimate $\hat{\beta}$ with the maximum likelihood method. Consequently the naive predictor $\hat{y}_i^*$ for $y_i$ is given by $\hat{y}_i^* = X_i \hat{\beta}$. The LP $\hat{y}_i^*$ for $y_i$ is given by solving (3). Since the parameter space for $y$ is discrete, in principle, we have to evaluate the $L(y^*|X)$ over all the $2^{20}$ possible combination of $y^*$ which requires a tremendous computational work. However, $L(y^*|X)$ is approximately equivalent to Bayesian posterior predictive density with flat prior. Therefore, we use Bayesian posterior simulation with MCMC to compute $\hat{y}_i^*$. We implement the Bayesian computation of the logistic model with flat prior via MCMClogit function of MCMCpack library (Martin & Quinn 2006) in R and summarize the predictive density with 4000 MCMC samples. We replicate the whole experiment 1000 times and summarize the results with the following statistics; the average percentage of correct prediction and the average absolute error per unit of standard error. We also use the approximate standard error given in (9) to compute the absolute error per unit of standard error.

In terms of the percentage of correct prediction both the methods, NP and LP are similar. However, in terms of absolute error per unit of SE, LP outperforms the others though the difference in very slim. A comparison of their empirical cumulative distribution is given in Figure 1.
Figure 1 presents the empirical cumulative densities of the average absolute error per unit of standard error. The thick solid line represents approximate likelihood method with SE calculated by using (9), the broken line presents likelihood prediction using Bayesian method implemented through MCMC and the solid thin line represents Naive prediction. They do not look very different from each other with LP(MCMC) being the best in terms of its range. However, the flat tail of NP and LP approximation can be explained from their inability to handle extreme predicted values. Therefore, we conclude that as long as the predicted values are not very extreme, the NP along with its standard error calculated from (9) can be used. In other special cases, the simulation method is to be preferred.

5 Use package pred for computation

This paper accompanies with an R add-in package, *pred*, to facilitate the computation of the predictive inference. Currently the package only covers the prediction methods discussed above. In the following we present some examples of producing likelihood prediction using the package: *pred*.

As a demonstrative example, we use the aids data set from Davison & Hinkley (1997). The data set contains information on the number of AIDS reports in England and Wales between the second quarter of 1983 and the fourth quarter of 1992. The observations are cross classified according to the diagnostic period and reporting delay, in three months interval. The AIDS cases are not all reported immediately after the diagnosis. Sometimes, it takes as long as 14 quarters from the diagnosis before some AIDS cases are reported. The data set is also available in *boot* library. The problem is to predict the number of AIDS reports at previous and current quarters while the prediction takes place at the 4\textsuperscript{th} quarter of 1992.
We fit a Poisson model for count response, number of AIDS reports. Based on the fitted model we can calculate predicted values in R in the following way.

```r
>library(boot)
>data(aids)
>g1<-glm(y~factor(delay)+factor(year):factor(quarter),
+family=poisson,data=aids[aids$dud==0,])  ## dud=0 indicates complete data
>predict(g1,newdata=aids[aids$dud==1])

375 389 390 403 404 ... 1.650853590 0.603278659 1.576317870 -0.074387547 0.793893031 ... 566 567 568 569 570 ... 1.584843220 1.310625193 0.334686204 1.202966781 2.176005992
Warning message:
In predict.lm(object, newdata, se.fit, scale = 1, type = ifelse(type == :
prediction from a rank-deficient fit may be misleading
```

Above we note that R produces the fitted values as the predicted value and that it does not produce any measure of prediction uncertainty. To work out those with the pred library we do the following

```r
>library(pred)
>pred(g1,newdata=aids[aids$dud==1]),se.pred=T)

375 389 390 403 404 ... 1.650853590 0.603278659 1.576317870 -0.074387547 0.793893031 ... 566 567 568 569 570 ... 1.584843220 1.310625193 0.334686204 1.202966781 2.176005992
Standard error of the predicted values
375 389 390 403 404 ... 2.802649e+01 1.941082e+00 1.947709e+01 1.262302e+00 2.516129e+00 ... 566 567 568 569 570 ... 2.651991e+01 9.370777e+00 1.815760e+00 7.168442e+00 1.255263e+03
```

Note that the default prediction method for pred is the approximate point prediction (Section 3.3) which can be specified through an argument method="PtP" in the pred function. Currently, "PtP" and the approximate predictive simulation (Section 3.1; "SLP1") are the two available options for the method argument to work with the glm object class. Getting the point prediction for each of the missing values was not the aim of the data analysis, rather we want to predict the total number of AIDS patients in a particular quarter within the years 1990–1992. We demonstrate how that can be obtained by making prediction for the 4th quarter of 1992.

```r
>p1<-pred(g1,newdata=aids[aids$dud==1]),method="SLP1",nsim=10000)
>s1<-p1$predsim ### Simulated prediction, one column for each observation.
>s<-apply(s1[,92:105],1,sum)
> 67+quantile(s,c(.025,.975))
    2.5% 97.5%
   137.000 2038.075
```
Hence, we get a 95% prediction interval (137, 2038) for the fourth quarter of 1992 which is much wider than the bootstrap prediction interval (357, 537) reported in Davison and Hinkley (1997). The huge length of the likelihood prediction interval can be explained from the fact that in quarter 4 of 1992 we only have report (67 infection) of the current quarter while prediction is based on what would have been reported in the next 14 quarters. In matter of fact there were about 2597 AIDS reports in England and Wales in 1992 (AIDS/HIV Quarterly Surveillance Tables 2003). Thus a mean prediction of 600 AIDS reports for quarter 4 in 1992 sounds reasonable.

6 Prediction of portfolio credit defaults

In order to make a sound prediction of portfolio credit risk and the value at risk we need a good statistical model as well as a sound predictive technique. The modeling issues are even challenging and the credit risk model often turns out to be very complicated. A huge amount of work on them are available in the literature (see e.g. Hao, Alam & Carling (2010) for a review). Taking the modeling framework presented in Alam (2010a) we demonstrate how to predict the future credit risk in a statistically sound way.

Let \( y_{ikt} \) denote the state of default (1 = default, 0 = non-default) of loan \( i \) (\( i = 1, 2, \ldots, n_{ikt} \)) in industry \( k \) (\( k = 1, 2, \ldots, K \)) and at time \( t \) (\( t = 1, 2, \ldots, T \)). Following the framework of GLMM, the conditional probability of default, \( p_{ikt} = E(y_{ikt}|u_{kt}) \), is modelled as \( y_{ikt}|u_{kt} \sim Bin (1, p_{ikt}) \) and

\[
g(p_{ikt}) = X_{ikt}\beta + Z_{ikt}u_t \tag{10}
\]

where \( g(.) \) is called the link function and \( g(p_{ikt}) = \log \left( \frac{p_{ikt}}{1-p_{ikt}} \right) \) gives a logistic mixed models which is also used in McNeil and Wendin. We use the logistic link in this paper. The use of \( g(p_{ikt}) = \log (-\log (1 - p_{ikt})) \) gives a complementary log-log model (as used in Carling, Rönnegård & Roszbach (2004)) and \( g(p_{ikt}) = \Phi^{-1}(p_{ikt}) \) produces a probit mixed model. \( X_{ikt} \) is a row vector of loan, industry and time specific covariates where each column of it corresponds to a covariate, \( \beta \) is a column vector of fixed effects parameters, \( Z_{ikt} \) is the \( ikt^{th} \) row of the design matrix associated with the random (unobservable) effects \( u_t = (u_{1t}, u_{2t}, \ldots, u_{Kt})^{T} \) and \( u_t \)'s are iid random variables having some specific distribution, \( f(u) \). Conventionally, the random effects vector, \( u_t \), is assumed to have a multivariate normal distribution with a 0 mean vector and a covariance matrix, \( D_t \), i.e.

\[
f(u) = \frac{|D_t|^{-\frac{1}{2}}}{(2\pi)^{\frac{K}{2}}} \exp \left[ -\frac{1}{2}u^{T}D^{-1}u \right].
\]

It was found in Alam (2010a) that the dependence in \( u_t \) over time, \( t \), is not significant.

After the model parameters have been estimated a naive prediction (NP) of credit defaults of \( ik^{th} \) loan at time \( t' > T \) would be given as

\[
p(y_{ikt'}) = X_{ikt}\hat{\beta}
\]

\( > 67 + \text{median(s)} \)
\( [1] 437 \)
\( > 67 + \text{mean(s)} \)
\( [1] 600.3048 \)
where $X_{ikt}$ is the row vector of the covariates observed at time $t$. After observing $X_{ikt}$ it makes sense to model $y_{ikt}$ conditional on $X_{ikt}$. But in this case we do not observe all the future covariates $X_{ikt'}$ at time $t'$. There are some covariates which e.g. macro indices (output gap, slope of the yield curve), firm specific characteristics such as, future sales, future asset values etc. can only be observed at time $t'$. Therefore, we also examine the effect of stochastic future covariates on the prediction performance.

### 6.1 Data description

We use the data set of Carling et al. (2004) to demonstrate the usefulness of the predictive likelihood in credit defaults prediction. The data set contains information on the credit history of 100,926 small and medium size aktiebolag, the approximate Swedish equivalent of US corporations and UK limited companies, from two major Swedish banks between the 2\textsuperscript{nd} quarter of 1994 and the 2\textsuperscript{nd} quarter of 2000. Besides the bank data on credit status, the data set contains information on accounting variables of the borrower companies e.g. sales, assets, liabilities etc., different remarks of the credit bureau and two macro variables being the slope of the yield curve and the output gap. We use the first 23 quarters’ data to build our model and the rest two quarters to evaluate the performance of the likelihood prediction compared to a naive frequentist’s prediction. This simulates a situation as to what would have happened if we were to predict credit defaults for two quarters ahead based on the information up to the 4\textsuperscript{th} quarter of 1999.

There are 6 clusters (industries) which are defined by merging the SNI industries (see www.scb.se) at the first two digits level, in a way that closely resembles the industry definition of Carling et al. (2004). We consider the credit defaults to be correlated both within and between clusters but not over time, as proposed in Alam (2010a). The mean part of the model is the same as in Carling et al. (2004) but we use a logit link whereas a complementary log-log link was used in Carling et. al (2004).

### 6.2 Data analysis and results

We fit the GLMM given in (10). In the implementation, we fit the GLMM with $u_{kt} \sim N(0, diag(\sigma_k^2))$. Later, we use the random effects obtained from the fitted model to estimate $\mathbf{D}$ i.e. we use

$$
\hat{\mathbf{D}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{u}_t \mathbf{u}_t^T
$$

where, $\mathbf{u}_t = (\hat{u}_{1t}, \hat{u}_{2t}, \ldots, \hat{u}_{Kt})^T$ is the vector of realized random effects for time $t$. The applicability of the above method was examined in (Alam & Carling 2008). In addition, we also fit a simple GLM having the same fixed effects as in (10) but the random effects are dropped.

In order to compare the performance of the GLMM and the GLM model in credit risk prediction, we use a simulation based prediction from the two models. We fit both of the models using the data up to quarter 23. Then we use them for predicting the credit defaults for quarter 24 and 25. Finally we compare the model implied defaults probability and the loss given defaults (LGD) defined as $LGD = EAD \times P(\text{Default})$ where $EAD$ is
the exposure (outstanding loan) at default with its real figures at the respective quarters. In simulation, we use the normal approximation method described in Section 3.1. For GLMM, we generate $\beta$ from $N\left(\hat{\beta}, I_\beta\right)$ where $\hat{\beta}$ and $I_\beta$ are the model estimated fixed effects and their respective observed information matrix. The random effects, $u_t$,’s, are simulated from $N\left(0, D\right)$. The predicted response is simulated, given the parameters and the covariates available at quarter 23 using a binomial model with mean $\mu_{ikt} = g^{-1}(\eta_{ikt})$ where $g$ is the logistic link function.

Table 1 Comparison between GLM and GLMM in credit defaults prediction

<table>
<thead>
<tr>
<th>Quarter</th>
<th>Probability of default (%)</th>
<th>Loss given default (% of total credit)</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>Observed 0.31 (0.20, 0.29)</td>
<td>GLM 0.25 (0.1, 0.48) GLMM 0.27 (0.01, 0.12)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLM 0.16 (0.1, 0.38) GLMM 0.22 (0.07, 0.11)</td>
</tr>
<tr>
<td>25</td>
<td>0.72 (0.11, 0.21)</td>
<td>0.22 (0.15, 0.40) GLMM 0.26 (0.09, 0.11)</td>
</tr>
<tr>
<td>24 &amp; 25</td>
<td>0.44 (0.18, 0.26)</td>
<td>0.22 (0.15, 0.40) GLMM 0.26 (0.09, 0.11)</td>
</tr>
</tbody>
</table>

Note: Values in the parenthesis represent 95% quantile interval

Table 1 presents the observed and the model implied probability of defaults and the loss given defaults per unit of outstanding total credits during the quarters 24 and 25. In application, to get the quarterly breakdown for quarter 24 and 25 by using all the time dependent variables fixed at quarter 23 while the time independent variables e.g. credit bureau remarks, credit duration are taken from the records at quarter 24 and 25 since they would be known at quarter 23. Therefore we see that the prediction model uses only the loans that were in the data base at quarter 23 and at least one of the other two quartets, 24 and 25. The observed proportion of defaults and the loss given default are calculated on the basis of all the loans which were actually available in the particular quarter.

Table 1 reveals that neither of the model performs very well, in terms of mean defaults prediction. For example at quarter 24, the observed proportion of defaults was 0.31% while the proportion was predicted as 0.25% by GLM and 0.27% by GLMM. A striking feature of GLM is that it produces the quantile interval misleadingly tight while the GLMM produces a wider quantile interval due to the consideration of the additional industry shocks (random effects). The 95% quantile interval of the GLMM almost captured the true proportion of defaults at quarter 24 but GLM did not. However, in quarter 25 both of the model failed by a big margin. In matter of fact, quarter 25 is an abnormal quarter in the historical credit risk data in that it has the highest proportion of default among the preceding 10 quarters. The overall proportion of defaults was only 0.55% during the whole 25 quarters while it was 0.19% in quarter 23 and 0.72% in quarter 25. Thus we conclude that none of the models are very successful to capture any abnormal sudden shock which might take place in the future two quarters.

A major drawback of both of the above model is that they consider all the covariates for quarter 24 and 25 as known at quarter 23 which is unrealistic. Though the accounting
variables may be treated as the same, since they are reported only once a year, the macro variables, output gaps and the slope of the yield curve are reported quarterly. The specific drawback of the GLMM is that it considers \(D\) as known while it uses an estimated \(D\) for prediction. Hence, an adjustment for the uncertainty in \(D\) is to be incorporated in the predictive simulation.

Since an estimate of the standard error of \(D\) is not readily available we use the Bayesian MCMC method to simulate \(D\) from its posterior distribution. In the application, we carry out the MCMC simulation in OpenBugs 3.0.2 (Spiegenhalter et al. 2007). It was however found that the Bugs software is very slow with full data set. Therefore, we use a subset of the data containing all the defaults and only a sample of equal number of non-defaults to get the posterior simulation from the GLMM model. We also use weakly informative prior, in stead of flat prior for the stability of the simulation. A detailed description of the Bayesian model is given in Alam (2010a).

We model the time series data on yield and output gap using an AR(2) and an AR(4) model and use the respective simulated forecasted values for quarter 24 and 25 for the predictive simulation of defaults probability and LGD. A time series plot of the output gap and the slope of yield curve along with their sample autocorrelation and sample partial autocorrelation plots are given in the appendix. The AIC suggests AR of order 2 for the slope of the yield curve and 7 for the output gap. However, to keep the model reasonably simple, from visual inspection of the time series properties of the above variables (see Appendix) we choose the lag orders 2 and 4, respectively. The simulated probability of defaults and the loss given defaults produced by GLMM after incorporating the additional uncertainties due to the macro variables (TS. forcast) and in \(D\) (D and TS) are given in Table 2.

<table>
<thead>
<tr>
<th>Quarter</th>
<th>Probability of default (%)</th>
<th>Loss given default (% of total credit)</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>Observed</td>
<td>TS. forecast</td>
</tr>
<tr>
<td>24</td>
<td>0.31</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>(0.10, 0.50)</td>
<td>(0.08, 0.52)</td>
</tr>
<tr>
<td>25</td>
<td>0.70</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>(0.07, 0.42)</td>
<td>(0.06, 0.45)</td>
</tr>
<tr>
<td>24 &amp; 25</td>
<td>0.44</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>(0.11, 0.43)</td>
<td>(0.10, 0.45)</td>
</tr>
</tbody>
</table>

Note: Values in the parenthesis represent 95% quantile interval

Table 2 shows that after correcting for the uncertainties involved with the future values of the macro variables and with the estimation of the covariance matrix, \(D\), the mean of the predicted proportion of defaults and loss given defaults do not change a lot but the prediction interval gets wider. For example, for combined credits history data on quarter 24 and 25, the GLMM predicted probability of defaults, before the correction, was 0.26% (see Table 1). After correcting for the macro variables’ uncertainty it becomes 0.24% (Table 2, 5th row and 3rd column) and after correcting for the uncertainties in both for the macro variables and in \(D\) it becomes 0.23% (Table 2, column "D and TS", row 5). The model generated 95% prediction interval (0.10, 0.45) now covers the actual
proportion of defaults observed in quarters 24 and 25 together (0.44%). However, the prediction interval for the loss given defaults, once again, fails to capture the abnormality of quarter 25.

7 Concluding discussion

Besides the likelihood prediction, there are other frequentist suggestion for prediction with linear and generalized linear mixed models. A major draw-back of those methods are that they should be assesses case by case and computation procedures are substantially different between different classes of models. However, the likelihood prediction offers a unified principle and method for making predictive inference in any situation.

The likelihood prediction may turn out to be computationally challenging but the simulation based methods can be applied in those cases. Applying the approximate predictive simulation for AIDS data and MCMC methods in predicting credit defaults risk we show that likelihood prediction may be useful in predicting future events with GLM and GLMM.

References


Alam, M. M. (2010b), Prediction with generalized linear and mixed models under covariate uncertainty, Unpublished manuscript, Dalarna University.


**URL:** [http://www.R-project.org](http://www.R-project.org)


**URL:** [http://mathstat.helsinki.fi/openbugs/Manuals/Manual.html](http://mathstat.helsinki.fi/openbugs/Manuals/Manual.html)

A Appendix: Time series properties of the macro variables

Fig. A. 1 Plot of the time series data on output gap and the slope of the yield curve.

Fig. A. 2 Sample autocorrelation and partial autocorrelation plot of output gap series.
A.1 Summary of the fitted time series models

<table>
<thead>
<tr>
<th>Time Series</th>
<th>Mean parameter</th>
<th>Lag 1</th>
<th>Lag 2</th>
<th>Lag 3</th>
<th>Lag 4</th>
<th>Variance Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output Gap</td>
<td>-1.13</td>
<td>0.92</td>
<td>0.40</td>
<td>-0.12</td>
<td>-0.25</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.09)</td>
<td>(0.13)</td>
<td>(0.12)</td>
<td>(0.09)</td>
<td></td>
</tr>
<tr>
<td>Slope of Yield Curve</td>
<td>0.58</td>
<td>1.17</td>
<td>-0.37</td>
<td>-</td>
<td>-</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>(0.09)</td>
<td>(0.08)</td>
<td>(0.08)</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Note: Values within parentheses present standard error

A.2 Summary of the time series model forecast

<table>
<thead>
<tr>
<th>Time Series</th>
<th>Quarter 24</th>
<th></th>
<th>Quarter 25</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed</td>
<td>Predicted</td>
<td>Observed</td>
<td>Predicted</td>
</tr>
<tr>
<td>Output Gap</td>
<td>0.14</td>
<td>0.80</td>
<td>0.56</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td>(0.54)</td>
<td>(0.73)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slope of Yield Curve</td>
<td>1.23</td>
<td>1.03</td>
<td>1.17</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>(0.94)</td>
<td>(1.45)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Values within parentheses present standard error

A.3 Available methods for predict function in R

The R function `methods(generic="name")` returns the list of the available methods associated with any generic function. Including the packages MASS, lme4, nlme and ghlm we obtain the following list of methods for the generic function `predict` in R.
> library(MASS)
> library(nlme)
> library(lme4)
> library(hglm)
> methods("predict")

[1] predict.ar* predict.Arima*
[3] predict.arima0* predict.glm
[5] predict.glmmPQL* predict.gls*
[9] predict.lda* predict.lm
[11] predict.lme* predict.lmList*
[13] predict.loess* predict.lqs*
[15] predict.mca* predict.mlm
[17] predict.nlme* predict.nls*
[19] predict.polr* predict.poly
[21] predict.ppr* predict.prcomp*
[23] predict.princomp* predict.qda*
[25] predict.rlm* predict.smooth.spline*
[27] predict.smooth.spline.fit* predict.StructTS*

Non-visible functions are asterisked