Robit Regression: A Simple Robust Alternative to Logistic and Probit Regression

Chuanhai Liu

Bell Laboratories, Lucent Technologies

E-mail: liu@research.bell-labs.com

March 7, 2006

Abstract

Logistic and probit regression models are commonly used in practice to analyze binary response data, but the maximum likelihood estimators of these models are not robust to outliers. This paper considers a robit regression model, which replaces the normal distribution in the probit regression model with a **t**-distribution with a known or unknown number of degrees of freedom. It is shown that (*i*) the maximum likelihood estimators of the robit model with a known number of degrees of freedom are robust; (*ii*) the robit link with about seven degrees of freedom provides an excellent approximation to the logistic link; and (*iii*) the robit link with a large number of degrees of freedom approximates the probit link. The maximum likelihood estimates can be obtained using efficient EM-type algorithms. EM-type algorithms also provide information that can be used to identify outliers, to which the maximum likelihood estimates of the logistic and probit regression coefficient would be sensitive. The EM algorithms for robit regression are easily modified to obtain efficient Data Augmentation (DA) algorithms for Bayesian inference with the robit regression model. The DA algorithms for robit regression model are much simpler to implement than the existing Gibbs sampler for the logistic regression model. A numerical example illustrates the methodology.

Key Words: Bayesian methods; the EM algorithm; the tobit model; Markov chain Monte Carlo; the PX-EM algorithm.

1 Introduction

The logistic and probit regression models are commonly used in practice to analyze binary response data, but many authors (see, Pregibon (1982) and the references therein) have shown that their maximum likelihood estimators are not robust. This paper considers a robit regression, which replaces the normal distribution in probit regression with a **t**-distribution with known or unknown degrees of freedom. The use of the **t**-distribution for robust estimation in the different context where the response variables are typically modeled with the normal distribution has been addressed by many authors (*e.g.*, Rubin, 1983; Lange, Little, and Taylor 1989; Liu and Rubin, 1995). As an alternative to logistic regression, this model has been previously suggested in the literature by Mudholkar and George (1978) and Albert and Chib (1993). Mudholkar and George (1978) discovered that a **t**-distribution with 9 degrees of freedom has the same kurtosis as the logistic regression. Albert and Chib (1993) suggested the use of a **t**-distribution with 8 degrees of freedom and provided the detailed implementation of the Gibbs sampler for Bayesian estimation.

It is shown that (*i*) the maximum likelihood estimators are robust if the number of degrees of freedom is known; (*ii*) the robit regression model with about seven degrees of freedom provides an excellent approximation to the logistic regression model; and (*iii*) the robit regression model with a large number of degrees of freedom approximates the probit regression model. Thus, in a certain sense, the robit regression model provides a rich class of models, including logistic and probit regression models as special cases, for analysis of binary response data.

This paper also provides efficient EM-type algorithms (Dempster, Laird, and Rubin, 1977; Liu, Rubin, and Wu, 1998) for finding the maximum likelihood estimates of the regression coefficients in the robit model. These algorithms provide information that can be used to identify outliers with too much influence on the maximum likelihood estimates of the regression coefficient under the logistic and probit models. Efficient Data-Augmentation (DA) algorithm (Tanner and Wong, 1987; Liu and Wu, 1999; Liu 1999) can be used to obtain estimates under a Bayesian robit model. The DA algorithm for the robit regression model is much simpler to implement than the existing Gibbs sampler (see, for example, Zeger and Karim (1991)) for the logistic regression model. Furthermore, the efficient DA algorithm can be extended to handle multivariate binary responses, as discussed briefly in end of the paper.

The rest of the paper is arranged as follows. Section 2 describes the robit model and its rela-

tionship with the probit and logistic models. Section 3 shows that the robust maximum likelihood estimators of the regression coefficients are robust. Section 4 formulates a complete data model for robit regression that can be used for maximum likelihood estimation using EM-type algorithms and for identifying outliers under logistic and probit models. Section 5 provides detailed implementation of the EM, ECME, and PX-EM algorithm for maximum likelihood estimation of the robit model. Section 6 describes the DA algorithms for fitting a Bayesian robit model. Section 7 illustrates the methodology with an example. Finally, Section 8 concludes with a few remarks.

2 The Robit Model

2.1 The logistic and probit models

Suppose that the observed data consist of n independent observations $\{(x_i, y_i) : i = 1, ..., n\}$ with a p-dimensional covariate vector x_i and binary response y_i that is either 0 or 1. The logistic regression model is specified by

$$\text{logit}\left(\text{pr}(y_i = 1|x_i, \beta)\right) = \ln \frac{\text{pr}(y_i = 1|x_i, \beta)}{1 - \text{pr}(y_i = 1|x_i, \beta)} = x_i'\beta \qquad (i = 1, ..., n). \tag{1}$$

The logistic regression model can also be derived by assuming that there are latent variables $z_i = x_i'\beta + e_i$, where e_i *logistic* with distribution function

$$F_{\text{logistic}}(x) = \frac{\exp\{x\}}{1 + \exp\{x\}} \tag{2}$$

and density

$$f_{\text{logistic}}(x) = \frac{\exp\{x\}}{(1 + \exp\{x\})^2} = \frac{1}{(\exp\{-x/2\} + \exp\{x/2\})^2} = \frac{1}{2(1 + \cosh(x))} \tag{3}$$

and y_i is one if $z_i > 0$ and zero otherwise. Then, the logistic regression model (1) is obtained as the marginal distribution of y_i. The maximum likelihood estimates of β can be obtained using the iterative re-weighted least-squares.

The probit model (*e.g.*, Albert and Chib, 1993), for which

$$\text{pr}(y_i = 1|x_i, \beta) = 1 - \text{pr}(y_i = 0|x_i, \beta) = \Phi(x_i'\beta) \qquad (i = 1, ..., n),$$

is obtained by replacing the logistic distribution for the latent error terms e_i with the standard normal distribution, where $\phi(x)$ and $\Phi(x)$ are the density and distribution functions of the standard

normal distribution, respectively. The maximum likelihood estimates of β in the probit model can be obtained using the EM algorithm (Dempster, Laird, and Rubin, 1977) or the PX-EM algorithm (Liu, Rubin, and Wu, 1998).

2.2 The robit model: a simple extension of the probit model

To have a robust model, following Lange, Little, and Taylor (1989), who replaced the normal distribution in linear regression model with a **t**-distribution to obtain robust estimators of linear regression coefficients, replace the normal distribution in probit regression model with the **t**-distribution with ν number of degrees of freedom. For computational simplicity, which itself is important in the current state of the art in statistics as discussed by Liu (2000), Albert and Chib (1993) suggested the use of a **t**-distribution with 8 degrees of freedom and provided the detailed implementation of the Gibbs sampler for Bayesian estimation.

We call this model *robit* regression, and denote by robit(ν) the robit regression model with ν degrees of freedom. More formally, the robit regression model for the data $\{(x_i, y_i) : i = 1, ..., n\}$ is

$$\text{pr}(y_i = 1|x_i, \beta) = 1 - \text{pr}(y_i = 0|x_i, \beta) = F_\nu(x_i'\beta) \qquad (i = 1, ..., n),$$

where $F_\nu(x)$ denotes the cdf of the **t** random variable with center zero, scale parameter one, and ν degrees of freedom. $F_\nu(x)$ has the density function

$$f_\nu(x) \equiv \frac{\Gamma((\nu+1)/2)}{(\pi\nu)^{1/2}\Gamma(\nu/2)(1 + x^2/\nu)^{(\nu+1)/2}} \qquad (x \in (-\infty, \infty)).$$

As $\nu \to \infty$, the robit(ν) model becomes the probit regression model.

2.3 The robit regression model with seven degrees of freedom: an approximation to the logistic model

Empirically, the robit link with about seven degrees of freedom approximates the logistic link, as Figure 1 suggests (The scale parameter $\sigma = 1.5484$ in Figure 1 was chosen by numerically minimizing $\max_{x_i}\{|F_\nu(x_i/\sigma) - F_{\text{logistic}}(x_i)| : x_i = -10 + 0.002i, i = 1, ..., 1000\}$ over σ. For $\sigma = 1.5484$, the maximum distance is about 0.0006). The quantiles below the 0.01 and 0.99 quantiles swing away from the reference line (dotted diagonal line), suggesting that the tail probabilities of the robit regression model are heavier than those of the logistic distribution. It is this tail property

that distinguishes the robit and logistic links in terms of robust estimation. To balance robustness and approximation to the logistic model, one may like to use the **t**-distribution with even smaller number of degrees of freedom, such as 6 and 5.

3 Robustness of Likelihood-based Inference Using Logistic, Probit, and Robit Regression Models

Consider the effects of a potential observation (x, y) on the estimates of $\text{pr}(y_i|x_i, \beta)$ for all i, or on the estimate of the regression coefficient vector β and consider the effective sample size s $(s > 0)$ of the potential observation. Without loss of generality, take $y = 1$. Let s $(s > 0)$ be the effective sample size. Denote by $\hat{\beta}_{+(x,y),s}$ the ML estimate of β with (y, x) included, that is,

$$\hat{\beta}_{+(x,y),s} = \arg\max_{\beta}\{\ell_{+(x,y),s}(\beta) \equiv \ell(\beta) + s\ln(\text{pr}(y|x, \beta))\},$$

where $\ell(\beta)$ denotes the log-likelihood given the observed data. If the ML estimates $\hat{\beta}$ and $\hat{\beta}_{+(x,y),s}$ are unique and finite, the *potential* influence of (x, y) is defined as

$$I(x, y) \equiv \lim_{s\to+0} \frac{\hat{\beta}_{+(x,y),s} - \hat{\beta}}{s}. \tag{4}$$

If the Hessian matrix $H(\hat{\beta}) = \partial^2\ell(\hat{\beta})/(\partial\beta\partial\beta')$ is negative definite, then

$$I(x, y) = -H^{-1}(\hat{\beta})\frac{\partial \ln \text{pr}(y|x, \hat{\beta})}{\partial\beta}.$$

Given the observed-data, $H(\hat{\beta})$ is fixed and can be viewed as a scaling matrix for the factor $\partial \ln \text{pr}(y|x, \hat{\beta})/\partial\beta$. Given the observed-data, $\hat{\beta}$ is also constant. To avoid the trivial cases, assume that all the components of $\hat{\beta}$ are non-zero so

$$\hat{\beta}'\frac{\partial \ln \text{pr}(y|x, \hat{\beta})}{\partial\beta}$$

is a convenient scalar factor. For the logistic regression model, $\hat{\beta}'\partial \ln \text{pr}(y|x, \hat{\beta})/\partial\beta = x'\hat{\beta}/(1+e^{x'\hat{\beta}})$, implying that the influence can be unbounded. For the probit regression model,

$$\frac{\partial \ln \text{pr}(y|x, \hat{\beta})}{\partial\beta} = \frac{\phi(x'\hat{\beta})}{\Phi(x'\hat{\beta})}x'\hat{\beta}.$$

When $x'\hat{\beta} \to -\infty$, this factor is approximately $-(x'\hat{\beta})^2$. This quadratic function in x indicates that the influence of (y, x) is unbounded and is more extreme than the influence under the logistic regression model.

For the robit regression model,

$$\frac{\partial \ln \mathrm{pr}(y|x,\hat{\beta})}{\partial \beta} = \frac{f_\nu(x'\hat{\beta})}{F_\nu(x'\hat{\beta})} x'\hat{\beta}.$$

This factor is bounded, and thereby the $I(x, y)$ is bounded because

$$\lim_{x'\hat{\beta}\to-\infty} \frac{f_\nu(x'\hat{\beta})}{F_\nu(x'\hat{\beta})} x'\hat{\beta} = \lim_{u\to-\infty} \frac{f_\nu(u)}{F_\nu(u)} u = 1 - \lim_{u\to-\infty} \frac{(\nu+1)u}{\nu+u^2} u = -\nu$$

and

$$\lim_{x'\hat{\beta}\to\infty} \frac{f_\nu(x'\hat{\beta})}{F_\nu(x'\hat{\beta})} x'\hat{\beta} = \lim_{\mu\to\infty} \frac{f_\nu(\mu)}{F_\nu(\mu)} \mu = 0.$$

4 Complete Data for Simple Maximum Likelihood Estimation

Let y_i denote the univariate binary response of the i-th individual, and let x_i denote the p-dimensional vector of covariates for $i = 1, ..., n$. Let

$$\tau_i|\theta \sim \mathrm{Gamma}(\nu/2, \nu/2) \qquad \text{and} \qquad z_i|(\tau_i, \theta) \sim \mathrm{N}(x_i'\beta, 1/\tau_i) \qquad (i = 1, ..., n),$$

where $\theta = (\beta, \nu)$ with β being the p-dimensional vector of regression coefficients and ν being the number of degrees of freedom. In the literature, τ_i is called weight, for example, in the context of iterative re-weighted least-squares. Then the robit regression model is completed by specifying

$$y_i = \begin{cases} 1, & \text{if } z_i > 0; \\ 0, & \text{if } z_i \le 0. \end{cases} \tag{5}$$

This complete-data model belongs to the exponential family. The sufficient statistics for θ are

$$S_\tau = \sum_{i=1}^n \tau_i,\ S_{\tau xx} = \sum_{i=1}^n \tau_i x_i x_i',\ S_{\tau zz} = \sum_{i=1}^n \tau_i z_i^2,\ S_{\tau xz} = \sum_{i=1}^n \tau_i x_i z_i,\ \text{and } S_{\ln\tau-\tau} = \sum_{i=1}^n (\ln\tau_i - \tau_i); \tag{6}$$

and the complete-data maximum likelihood estimate of $\theta = (\beta, \nu)$ is given by

$$\hat{\beta} = S_{\tau xx}^{-1} S_{\tau xz} \quad \text{and} \quad \hat{\nu} = \arg\max_\nu \left[-n \ln\Gamma(\nu/2) + n(\nu/2)\ln(\nu/2) + (\nu/2) S_{\ln\tau-\tau}\right].$$

Let $\mu_i = x_i'\beta$, denote by t_ν the t-deviate with location zero, scale parameter one, and the number of degrees of freedom ν, and denote by $f_{t_\nu}(.)$ the probability density of t_ν, *i.e.*,

$$f_{t_\nu}(z) = c_\nu(1 + z^2/\nu)^{-(\nu+1)/2}$$

with the normalizing constant

$$c_\nu = (\pi\nu)^{-1/2}\Gamma((\nu+1)/2)\Gamma^{-1}(\nu/2).$$

Then

$$\begin{aligned}
\hat{\tau}_i \equiv \mathrm{E}(\tau_i|Y_{\mathrm{obs}},\theta) &= \mathrm{E}(\mathrm{E}(\tau_i|z_i,Y_{\mathrm{obs}},\theta)) = \mathrm{E}\left(\frac{1+1/\nu}{1+(z_i-\mu_i)^2/\nu}\middle| Y_{\mathrm{obs}},\theta\right) \\
&= \frac{\nu+1}{\nu}\frac{\int_{\{z:I(z\geq-\mu_i)=y_i\}} c_\nu(1+z^2/\nu)^{-(\nu+3)/2}dz}{\int_{\{z:I(z\geq-\mu_i)=y_i\}} f_{t_\nu}(z)dz} \\
&= \begin{cases} \frac{\mathrm{pr}(t_{\nu+2}<-(1+2/\nu)^{1/2}\mu_i)}{\mathrm{pr}(t_\nu<-\mu_i)}, & \text{if } y_i=0; \\ \frac{\mathrm{pr}(t_{\nu+2}>-(1+2/\nu)^{1/2}\mu_i)}{\mathrm{pr}(t_\nu>-\mu_i)}, & \text{if } y_i=1 \end{cases} \\
&= \frac{y_i-(2y_i-1)\mathrm{pr}(t_{\nu+2}<-(1+2/\nu)^{1/2}\mu_i)}{y_i-(2y_i-1)\mathrm{pr}(t_\nu<-\mu_i)}, \qquad (7) \\
\mathrm{E}(\tau_i(z_i-\mu_i)|Y_{\mathrm{obs}},\theta) &= \frac{\nu+1}{\nu}\mathrm{E}\left(\frac{z_i-\mu_i}{1+(z_i-\mu_i)^2/\nu}\middle| Y_{\mathrm{obs}},\theta\right) \\
&= \frac{\nu+1}{\nu}\frac{\int_{\{z:I(z\geq-\mu_i)=y_i\}} c_\nu z(1+z^2/\nu)^{-(\nu+3)/2}dz}{\int_{\{z:I(z\geq-\mu_i)=y_i\}} f_{t_\nu}(z)dz} \\
&= \frac{(2y_i-1)f_{t_\nu}(\mu_i)}{y_i-(2y_i-1)\mathrm{pr}(t_\nu<-\mu_i)} \\
&= \hat{\tau}_i\frac{(2y_i-1)f_{t_\nu}(\mu_i)}{y_i-(2y_i-1)\mathrm{pr}(t_{\nu+2}<-(1+2/\nu)^{1/2}\mu_i)}, \\
\mathrm{E}(\tau_i(z_i-\mu_i)^2|Y_{\mathrm{obs}},\theta) &= \frac{\nu+1}{\nu}\mathrm{E}\left(\frac{(z_i-\mu_i)^2}{1+(z_i-\mu_i)^2/\nu}\middle| Y_{\mathrm{obs}},\theta\right) \\
&= (\nu+1)\frac{\int_{\{z:I(z\geq-\mu_i)=y_i\}} c_\nu(z^2/\nu)(1+z^2/\nu)^{-(\nu+3)/2}dz}{\int_{\{z:I(z\geq-\mu_i)=y_i\}} f_{t_\nu}(z)dz} \\
&= \nu+1-\nu\hat{\tau}_i.
\end{aligned}$$

With

$$\hat{z}_i \equiv \mu_i + \frac{(2y_i-1)f_{t_\nu}(\mu_i)}{y_i-(2y_i-1)\mathrm{pr}(t_{\nu+2}<-(1+2/\nu)^{1/2}\mu_i)}, \qquad (8)$$

it follows then

$$\mathrm{E}(\tau_i z_i|Y_{\mathrm{obs}},\theta) = \mathrm{E}(\tau_i(z_i-\mu_i)|Y_{\mathrm{obs}},\theta) + \mu_i\mathrm{E}(\tau_i|Y_{\mathrm{obs}},\theta) = \hat{\tau}_i\hat{z}_i,$$

and

$$\begin{aligned}\mathrm{E}(\tau_i z_i^2|Y_{\mathrm{obs}},\theta) &= \mathrm{E}(\tau_i(z_i-\mu_i)^2|Y_{\mathrm{obs}},\theta)+2\mu_i\mathrm{E}(\tau_i(z_i-\mu_i)|Y_{\mathrm{obs}},\theta)+\mu_i^2\mathrm{E}(\tau_i|Y_{\mathrm{obs}},\theta)\\ &= \nu+1-\nu\hat{\tau}_i+\hat{\tau}_i\left[\mu_i^2+2\mu_i(\hat{z}_i-\mu_i)\right]. \qquad (9)\end{aligned}$$

When the conditional expectation of the sufficient statistics is calculated at the ML estimate of θ,

$$\hat{\beta}=\left(\sum_{i=1}^n\hat{\tau}_i x_i x_i'\right)^{-1}\left(\sum_{i=1}^n\hat{\tau}_i x_i\hat{z}_i\right),$$

which is the ML estimate of β in the linear regression $\hat{z}_i\sim N(x_i'\beta,\hat{\tau}_i)$.

Letting $\nu\to\infty$ gives the complete-data probit regression model and the conditional expectations of the associated sufficient statistics:

$$\lim_{\nu\to\infty}\hat{\tau}_i=1,\quad \lim_{\nu\to\infty}\hat{z}_i=\mu_i+\frac{(2y_i-1)\phi(\mu_i)}{y_i-(2y_i-1)\Phi(-\mu_i)},\quad \text{and}\quad \lim_{\nu\to\infty}E(z_i^2|Y_{\mathrm{obs}},\theta)=1+\mu_i\hat{z}_i.$$

The last equality is obtained using the fact that $\nu+1-\nu\hat{\tau}_i\to 1-\mu_i z_i+\mu_i^2$ as $\nu\to\infty$.

5 Maximum Likelihood Estimation Using EM-type Algorithms

5.1 MLE of the regression coefficients β with known number of degrees of freedom ν using EM

With the complete-data $\{(x_i,y_i,z_i,\tau_i):i=1,...,n\}$ described in Section 4, the EM algorithm for finding the MLE of β with known ν is as follows. At iteration $t+1$ with input $\beta^{(t)}$,

E-step of EM: Compute $\hat{\tau}_i$ and $\hat{z}_i$ for all $i=1,...,n$ in (7) and (8) with $\theta=(\beta^{(t)},\nu)$, and then the expected sufficient statistics $\hat{S}_{\tau xx}=\sum_{i=1}^n\hat{\tau}_i x_i x_i'$ and $\hat{S}_{\tau xy}=\sum_{i=1}^n\hat{\tau}_i x_i\hat{z}_i$.

M-step of EM: Update β: $\beta^{(t+1)}=\hat{S}_{\tau xx}^{-1}\hat{S}_{\tau xy}$.

5.2 MLE of $\theta=(\beta,\nu)$ with unknown number of degrees of freedom ν using ECME

To use the EM algorithm to find the MLE of $\theta=(\beta,\nu)$ when the number of degrees of freedom ν is unknown, compute

$$\mathrm{E}((\ln\tau_i-\tau_i)|Y_{\mathrm{obs}},\theta)=\phi((\nu+1)/2)-\ln((\nu+1)/2)+\mathrm{E}\left(\ln\frac{\nu+1}{\nu+(z_i-\mu_i)^2}\middle|Y_{\mathrm{obs}},\theta\right)-\hat{\tau}_i \qquad (10)$$

for all $i = 1, ..., n$, where $\phi(\alpha) \equiv d \ln(\Gamma(\alpha))/d\alpha = \Gamma'(\alpha)/\Gamma(\alpha)$ is the digamma function. Because there are no (obvious) numerical methods for computing the conditional expectation term in (10) and ECME typically converges dramatically faster than EM, we use ECME with two constrained maximization (CM) steps: one CM step maximizes the expected complete-data log-likelihood over β with ν fixed at its current estimate; and the other CM step maximizes the constrained actual likelihood over ν with β fixed at its current estimate, where the constrained likelihood function of ν given β is

$$\ell(\nu|\beta, Y_{\text{obs}}) = \sum_{i=1}^{n} \ln\left(y_i(1 - \text{pr}(t_\nu < -\mu_i)) + (1 - y_i)\text{pr}(t_\nu < -\mu_i)\right). \tag{11}$$

The ECME algorithm for finding the MLE of $\theta = (\beta, \nu)$ is as follows. At iteration $t + 1$ with input $\theta^{(t)} = (\beta^{(t)}, \nu^{(t)})$,

E-step of ECME: The same as the E-step of EM: condition on the current parameter estimates, $\theta^{(t)} = (\beta^{(t)}, \nu^{(t)})$.

CM-step 1 of ECME: The same as the M-step of EM.

CM-step 2 of ECME: Search for the $\nu^{(t+1)}$ that maximizes $\ell(\nu|\beta^{(t+1)}, Y_{\text{obs}})$.

Then update ν using, for example, the half-interval method (Carnahan, Luther, and Wilks, 1969) to maximize $\ell(\nu|\beta, Y_{\text{obs}})$ in the likelihood function (11).

5.3 MLE of the robit model using PX-EM: a more efficient algorithm for computing $(\hat{\beta}, \hat{\nu})$

Liu, Rubin, and Wu (1998) show that the PX-EM algorithm, which makes use of the extra information captured in the imputed complete data, converges much faster than the EM algorithm for finding the MLE of the **t**-distribution and the probit regression model. Here PX-EM is used to find the MLE of the robit model, which involves both the **t**-distribution and the probit model. To make use of the extra information captured in the complete data, following Liu, Rubin, and Wu (1998), the complete-data model is extended as

$$(\tau_i/\alpha)|\theta^* \sim \text{Gamma}(\nu^*/2, \nu^*/2),\ z_i|(\tau_i, \theta^*) \sim \text{N}(x_i\beta^*, \sigma^2/\tau_i),$$

and

$$y_i = I(z_i \geq 0)$$

for $i = 1, ..., n$, where $\theta^* = (\beta^*, \nu^*, \alpha, \sigma)$ with $\alpha > 0$ and $\sigma > 0$. The observed-data model is preserved with the reduction function

$$\beta = (\alpha/\sigma)\beta^* \quad \text{and} \quad \nu = \nu^*. \tag{12}$$

The complete-data sufficient statistics for the expanded parameters θ^* are given in (6). The complete-data MLE of θ^* is given by

$$\hat{\alpha} = n^{-1} \sum_{i=1}^{n} \tau_i, \quad \hat{\sigma}^2 = n^{-1}(S_{\tau zz} - S'_{\tau xz} S^{-1}_{\tau xx} S_{\tau xz}),$$

and $\hat{\beta^*}$ and $\hat{\nu^*}$ are the same as $\hat{\beta}$ and $\hat{\nu}$, respectively. Compared to the EM algorithm in Section 5.1 and the ECME algorithm in Section 5.2, the corresponding PX-EM and PX-ECME algorithms require only simple extra computation, namely, the conditional expectations of S_τ and $S_{\tau zz}$. The PX-EM algorithm for finding the regression coefficients β with known number of degrees of freedom ν is then a simple extension of the EM algorithm and is given as follows. At iteration $t + 1$ with input $\beta^{(t)}$,

E-step of PX-EM: The same as the E-step of EM, except for the extra calculation of the conditional expectations $\hat{S}_\tau = \sum_{i=1}^n \hat{\tau}_i$ and $\hat{S}_{\tau zz} = n(\nu + 1) - \nu \sum_{i=1}^n \hat{\tau}_i + \sum_{i=1}^n \hat{\tau}_i(2\mu_i \hat{z}_i - \mu_i^2)$.

M-step of PX-EM: Compute $\hat{\beta^*} = \hat{S}^{-1}_{\tau xx} \hat{S}_{\tau xy}$, $\hat{\alpha} = n^{-1} \hat{S}_\tau$, and $\hat{\sigma}^2 = n^{-1}(\hat{S}_{\tau zz} - \hat{S}'_{\tau xz} \hat{S}^{-1}_{\tau xx} \hat{S}_{\tau xz})$ and then apply the reduction function to update β: $\beta^{(t+1)} = (\hat{\alpha}/\hat{\sigma})\hat{\beta^*}$

With unknown number of degrees of freedom ν, the ECME algorithm is then extended to the following PX-ECME algorithm. At iteration $t + 1$ with input $\theta^{(t)} = (\beta^{(t)}, \nu^{(t)})$,

E-step of PX-ECME: The same as the E-step of PX-EM, just conditioning on the parameter estimates, $\theta^{(t)} = (\beta^{(t)}, \nu^{(t)})$.

CM-step 1 of PX-ECME: The same as the M-step of PX-EM.

CM-step 2 of PX-ECME: The same as the CM-step 2 of ECME.

6 Bayesian Estimation of the Robit Regression Model with Known Number of Degrees of Freedom Using DA Algorithms

For Bayesian estimation of the robit regression model, this paper uses the multivariate **t**-distribution

$$\text{pr}(\beta) = \mathbf{t}_p(0, S_0^{-1}, \nu_0) \tag{13}$$

as the prior distribution for the regression coefficients β, where S_0 is a known $(p \times p)$ non-negative definite scatter matrix and ν_0 is the known degrees of freedom. When S_0 is positive definite, the posterior distribution of β is proper because the likelihood is bounded. When $S_0 = 0$ the prior distribution for β is flat and β may have an improper posterior in the sense that $\int_\beta \text{pr}(\beta)\ell(\beta|Y_{\text{obs}})d\beta = \infty$. Chen and Shao (1999) discuss this issue.

The **t**-distribution (13) can be represented as the marginal distribution of β in the following well-known hierarchical structure

$$\tau_0 \sim \text{Gamma}(\nu_0/2, \nu_0/2) \quad \text{and} \quad \beta|\tau_0 \sim \text{N}_p(0, S_0^{-1}/\tau_0). \tag{14}$$

Like the missing weights τ_i $(i = 1, ..., n)$, in the sequel τ_0 is treated as missing. Corresponding to the complete data augmented for implementation of the EM algorithms, the complete data for generating draws of β from its posterior distribution using the DA algorithm consist of Y_{obs}, $z = (z_1, ..., z_n)$ and $\tau = (\tau_0, \tau_1, ..., \tau_n)$.

6.1 Simulating the posterior of β using the DA algorithm

Similar to the implementation of the EM algorithm for finding the ML estimates of β, the implementation of the DA algorithm for simulating the posterior of β consists of an Imputation (I) step and a Posterior simulation (P) step, which are given as follows.

I-step of DA: Conditioning on the observed data and the current draw of β, draw $\{(z_i, \tau_i) : i = 1, ..., n\}$ by first taking a draw of z_i from the truncated $\mathbf{t}(\mu_i = x_i'\beta, 1, \nu)$, which is either left $(y_i = 1)$ or right $(y_i = 0)$ truncated at 0, and then taking a draw of τ_i from

$$\text{Gamma}\left(\frac{\nu+1}{2}, \frac{\nu + (z_i - \mu_i)^2}{2}\right)$$

for all $i = 1, ..., n$, and a draw of τ_0 from its distribution given in (14).

P-step of DA: Conditioning on the current draws of $\{(z_i, \tau_i) : i = 1, ..., n\}$, draw β from the p-variate normal distribution

$$\mathrm{N}_p\left(\hat{\beta}, (\tau_0 S_0 + S_{\tau xx})^{-1}\right),$$

where

$$\hat{\beta} = (\tau_0 S_0 + S_{\tau xx})^{-1} S_{\tau xz}, \tag{15}$$

and $S_{\tau xx}$ and $S_{\tau xz}$ are defined in (6).

6.2 Simulating the posterior of β via efficient DA algorithms

Like the EM algorithm, the DA algorithm can converge very slowly. The DA algorithm can be accelerated by using the ideas of the PX-EM algorithm. Two approaches, which are practically equivalent, can be taken. One is the PX-DA algorithm (Liu and Wu, 1999; see also Meng and van Dyk, 1999), which extends the PX-EM algorithm by making use of the group transformation indexed by the expanded parameters used in the PX-EM algorithm. Technically, what is needed is a prior on the group transformation. This prior specification can be avoided by using the CA-DA algorithm (Liu, 1999), which adjusts the current draws of the parameters and missing data by redrawing the sufficient statistics of the expanded parameters and the original parameters conditioning on their complements. Typically, the complements take the form of residuals, or more exactly, pivotal quantities. Here, we take the CA-DA approach.

First, adjust individual scores z_i for their common scale parameter σ. The sufficient statistic for σ, after integrating out the regression coefficients β, is

$$s^2 = \sum_{i=1}^{n} \tau_i \left(z_i - x_i'\hat{\beta}\right)^2 + \hat{\beta}'\tau_0 S_0 \hat{\beta},$$

where $\hat{\beta} = (\tau_0 S_0 + S_{\tau xx})^{-1} S_{\tau xz}$. To draw (s^2, β) with z_i $(i = 1, ..., n)$ fixed up to a proportionality constant (*i.e.*, the scale of z_is), take the re-scaling transformation

$$z_i^* = z_i / s \qquad (i = 1, ..., n). \tag{16}$$

with the constraint

$$\sum_{i=1}^{n} \tau_i \left(z_i^* - x_i'\hat{\beta}^*\right)^2 + (\hat{\beta}^*)'\tau_0 S_0 \hat{\beta}^* = 1, \tag{17}$$

where $\hat{\beta}^* = (\tau_0 S_0 + S_{\tau xx})^{-1} S_{\tau xz^*}$ with $S_{\tau xz^*}$ obtained from $S_{\tau xz}$ by substituting z_i^* for z_i. Since the transformation (16) from $(z*, s)$ to z with the constraint (17) is one-to-one, a version of the CA-DA algorithm can be obtained from DA by replacing the P-step of DA with a step that draws (β, s^2), conditioning on z^*. The Jacobean of the transformation from (z, β) onto $(z^*, s, \eta = \beta)$ with the constraints (17), as a function of (s, η), is proportional to s^{n-1}. The conditional distribution of (s, η) given z^* is then

$$\text{pr}(s, \eta | \tau, z^*, Y_{\text{obs}}) = \text{pr}(s | \tau, z^*, Y_{\text{obs}}) \cdot \text{pr}(\eta | s, \tau, z^*, Y_{\text{obs}}),$$

where $\text{pr}(s^2 | \tau, z^*, Y_{\text{obs}}) = \text{Gamma}(n/2, 1/2)$ and $\text{pr}(\eta | s, \tau, z^*, Y_{\text{obs}}) = \text{N}(s\hat{\beta}^*, (\tau_0 S_0 + S_{\tau xx})^{-1})$. This leads to the following efficient DA algorithm, denoted by E-DA 1,

I-step of E-DA 1: This is the same as the I-step of DA.

P-step of E-DA 1: This is the same as the P-step of DA, except for rescaling $\hat{\beta}$ by a factor of $\chi_n / \left[\sum_{i=1}^n \tau_i (z_i - x_i' \hat{\beta})^2 + \hat{\beta}' \tau_0 S_0 \hat{\beta}\right]^{1/2}$, where χ_n^2 is a draw from the chi-square distribution with n degrees of freedom.

For the probit regression model, *i.e.*, $\nu = \infty$ and thereby $\tau_i = 1$ for all $i = 1, ..., n$, E-DA 1 is equivalent to the PX-DA algorithm of Liu and Wu (1999), who considered a flat prior on β. The P-step of E-DA 1 implicitly integrates out the scale of $z_i s$, which explains intuitively why E-DA 1 converges faster than DA.

Second, adjust the individual weights for their scale to obtain a DA sampling scheme that is even faster than E-DA 1. Let

$$w = \sum_{i=0}^n \nu_i \tau_i \quad \text{and} \quad ws^2 = \sum_{i=1}^n \tau_i (z_i - x_i' \hat{\beta})^2 + \hat{\beta}' \tau_0 S_0 \hat{\beta},$$

where $\nu_i = \nu$ for all $i = 1, ..., n$. Take the transformation

$$\tau_i = w \tau_i^* \quad (i = 0, ..., n; w > 0) \quad \text{and} \quad z_i = s z_i^* \quad (i = 1, ..., n; w > 0)$$

with the constraints

$$\sum_{i=0}^n \nu_i \tau_i^* = 1 \qquad \text{and} \qquad \sum_{i=1}^n \tau_i^* (z_i^* - x_i' \hat{\beta}^*)^2 + \hat{\beta}' \tau_0 S_0 \hat{\beta} = 1, \tag{18}$$

where $\hat{\beta}^* = (\tau_0^* S_0 + S_{\tau^* xx})^{-1} S_{\tau^* xz^*} = (\tau_0 S_0 + S_{\tau xx})^{-1} S_{\tau xz^*}$ with $S_{\tau^* xx}$ and $S_{\tau^* xz^*}$ obtained from $S_{\tau xx}$ and $S_{\tau xz}$, respectively, by replacing τ_i with τ_i^* and z_i with z_i^*. The Jacobean of the transformation from (τ, z, β) to $(\tau^*, z^*, w, s, \eta = \beta)$ with the constraints (18), as a function of (w, s, η) is proportional to $w^n s^{n-1}$. Thus, conditioning on z^*, τ^*, and $Y_{\rm obs}$, $(w, s, \eta = \beta)$ is distributed as

$$\mathrm{pr}(w|z^*, \tau^*, Y_{\rm obs}) \cdot \mathrm{pr}(s|w, z^*, \tau^*, Y_{\rm obs}) \cdot \mathrm{pr}(\beta|w, s, z^*, \tau^*, Y_{\rm obs}),$$

where $\mathrm{pr}(w|z^*, \tau^*, Y_{\rm obs}) = \mathrm{Gamma}((\nu_0 + n\nu)/2, 1/2)$, $\mathrm{pr}(s^2|w, z^*, \tau^*, Y_{\rm obs}) = \mathrm{Gamma}(n/2,\ w/2)$, and $\mathrm{pr}(\beta|w, s, z^*, \tau^*, Y_{\rm obs}) = \mathrm{N}_p(s\hat{\beta}^*, w^{-1}(\tau_0^* S_0 + S_{\tau^* xx})^{-1})$. This leads to the following efficient DA algorithm, denoted by E-DA 2,

I-step of E-DA 2: This is the same as the I-step of DA.

P-step of E-DA 2: This is the same as the P-step of E-DA 1, except for rescaling the draw of β by a factor of $\left(\sum_{i=0}^n \nu_i \tau_i / \chi^2_{\nu_0 + n\nu}\right)^{1/2}$, where $\chi^2_{\nu_0+n\nu}$ is a draw from the chi-square distribution with $\nu_0 + n\nu$ degrees of freedom.

The P-step of E-DA 2 implicitly integrates out both the scale of z_is and the scale of τ_is, which explains why E-DA 2 is converges faster than both DA and E-DA 1.

7 A Numerical Example

The data are taken from Finney (1947) and consist of 39 binary responses denoting the presence ($y = 1$) or absence ($y = 0$) of vaso-constriction of the skin of the subjects after inspiration of a volume V of air at inspiration rate R. The data were obtained from repeated measurements on three individual subjects, the numbers of observations per subject being 9, 8, and 22. Finney (1947) found no evidence of inter-subject variability, treated the data as 39 independent observations, and analyzed the data using the probit regression model with V and R in the logarithm scale as covariates. This data set was also analyzed by Pregibon (1982), using robust procedures (called resistant fitting methods) as alternatives to logistic regression.

The data are displayed in Figure 2. The fitted probability contours obtained from the MLE indicate that there is little difference between the the fitted probit and logistic regression models. From these contours, the robit(7) and logistic models are almost identical, suggesting again that the robit(7) model is a nice alternative to the logistic model in the sense that the robit(7) regression

model provides results can be understood as those from the logistic model and that the MLE of robit(7) regression model is robust.

The EM algorithm was applied to choose the number of degrees of freedom. The algorithm was stopped when the likelihood increment becomes numerically instable because of the accuracy in evaluation of the probability functions of the **t**distributions. The estimate of $\hat{\nu}$ is about 0.11 with the likelihood value -10.62. The fitted robit models with various numbers of degrees of freedom are represented by the probability contours in Figure 3. The use of a small number of the degrees of freedom is intuitively suggested by the data, in which the observations with positive responses and those with negative responses can be almost separated by a line on the plane of $\ln(V)$ and $\ln(R)$ except for the three observations with $i = 4$, 18, and 24. These three observations are identified from the fitted individual weights. Pregibon (1982) also found that these three observations are influential to the ML estimation of the logistic model. The fitted 0.1, 0.5, and 0.9 contours by Pregibon are similar to those obtained from the robit model with about $\nu = 2$ degrees of freedom.

The Bayesian results using the prior distribution with $\nu_0 = 1$ and $S_0 = 0.0001I$, which is practically flat for the skin vaso-constriction data, were obtained using the DA algorithms. Figure 4 displays the posterior probability

$$\mathrm{pr}(y = 1|x) = \int_\beta \mathrm{pr}(y = 1|x, \beta) f(\beta|Y_{\mathrm{obs}}) d\beta$$

with various known numbers of degrees of freedom, where $f(\beta|Y_{\mathrm{obs}})$ is the posterior distribution of β. These results are similar to those obtained from the ML fitting. From Finney (1947), it is of interest to compare the difference $\beta_{\mathrm{RATE}} - \beta_{\mathrm{VOL}}$. Figure 5 shows the posterior distributions (in solid line) of the difference $\beta_{\mathrm{RATE}} - \beta_{\mathrm{VOL}}$ obtained from the robit model with $\nu = \infty, 7, 2$, or 1. The posterior probability $\mathrm{pr}(d > 0|Y_{\mathrm{obs}})$ increases from 0.68 to 0.91 as ν decrease to 1. Figure 5 also shows the corresponding results obtained with the two most influential observations ($i = 4$ and 8) removed. These results suggest that the robit model with a small number of degrees of freedom provides reliable inference, for example, regarding the difference between β_{RATE} and β_{VOL}.

8 Conclusion

It has been shown that the robit model is a useful robust alternative to the probit and logistic models for analyzing binary response data. The advantages of using the robit model include (1)

the inference based on the robit model is robust to the presence of outlying observations, and (2) computation for a Bayesian robit regression model using Markov chain Monte Carlo (MCMC) methods is simpler than that for the logistic model (see, for example, Zeger and Karim (1991)), especially when the model is extended to allow for random effects. Since robit(ν) with small ν gives more weight to the observations that are close to the dividing line ($\text{pr}(y = 1|x) = (\text{pr}(y = 0|x) = 1/2$ when they agree with the fitted model, the robit model with a small number of degrees of freedom should also be useful in classification. In addition, as with the probit model (*e.g.*, Albert and Chib, 1993; and Chib and Greenberg, 1998), the extension of the robit model to correlated multivariate responses is straightforward, where the efficient DA algorithms appear to be especially useful (Liu, 2000).

Acknowledgement

The author thanks Dr. Diane Lambert for her numerous insightful and constructive comments.

References

Carnahan, B., Luther, H., and Wilks, J. O. (1969). *Applied Numerical Methods*. John Wiley, New York.

Chen, M-H, and Shao, Q-M. (2000). Propriety of posterior distribution for dichotomous quantal response models, *Proceedings of the American Mathematical Society*, to appear.

Chib, S. and Albert, J. H. (1993). Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Assoc.* **88**, 669-679.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* **85**, 347-361.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B **39**, 1–38.

Finney, D. J. (1947). The estimation from individual records of the relationship between dose and quantal response. *Biometrika*, **34**, 320-334.

Hastie, T. J. and Pregibon, D. (1993). General Linear Models. *Statistical Models in S* (eds. Chambers, J. M. and Hastie, T. J.), 105-248, Chapman & Hall, New York.

Lange, K. L., Little, R. J. A., and Taylor, J. M. G. (1989), Robust Statistical Modeling Using the **t** Distribution, *Journal of the American Statistical Association*, **84**, 881-896.

Liu, C. (1999). Covariance adjustment for Markov chain Monte Carlo — a general framework. Technical report, Bell Labs.

Liu, C. (2000). Comment on "*The Art of Data Augmentation*" by Meng and van Dyk. *Journal of Computational and Graphical Statistics*, to Appear.

Liu, C., Rubin, D. B., and Wu, Y. N. (1998). Parameter expansion to accelerate EM: the PX-EM algorithm, *Biometrika* **85**, 755-770.

Liu, C and Rubin, D. B. (1995). ML Estimation of the Multivariate **t** Distribution With Unknown Degrees of Freedom, *Statistica Sinica*, **5**, 19-39.

Liu, J. S. and Wu, Y. (1999). Parameter expansion scheme for data augmentation. *J. Am. Statist. Assoc.* **94**, 1264-1274.

Meng, X. L., and van Dyk, D. A. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika* **86**, 301-320.

Mudholkar, G. S. and George E. O. (1978). A remark on the shape of the logistic distribution. *Biometrika*, **65**, 667-668.

Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrics*, **38**, 485-498.

Rubin, D. B. (1983). "Iteratively reweighted least squares" in Encyclopedia of Statistical Sciences, Vol. 4, John Wiley & Sons. *pp.* 272-275.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Am. Statist. Assoc.* **82**, 528–550.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects; A Gibbs sampling approach. *J. of American Assoc.* **86**, 79-86.

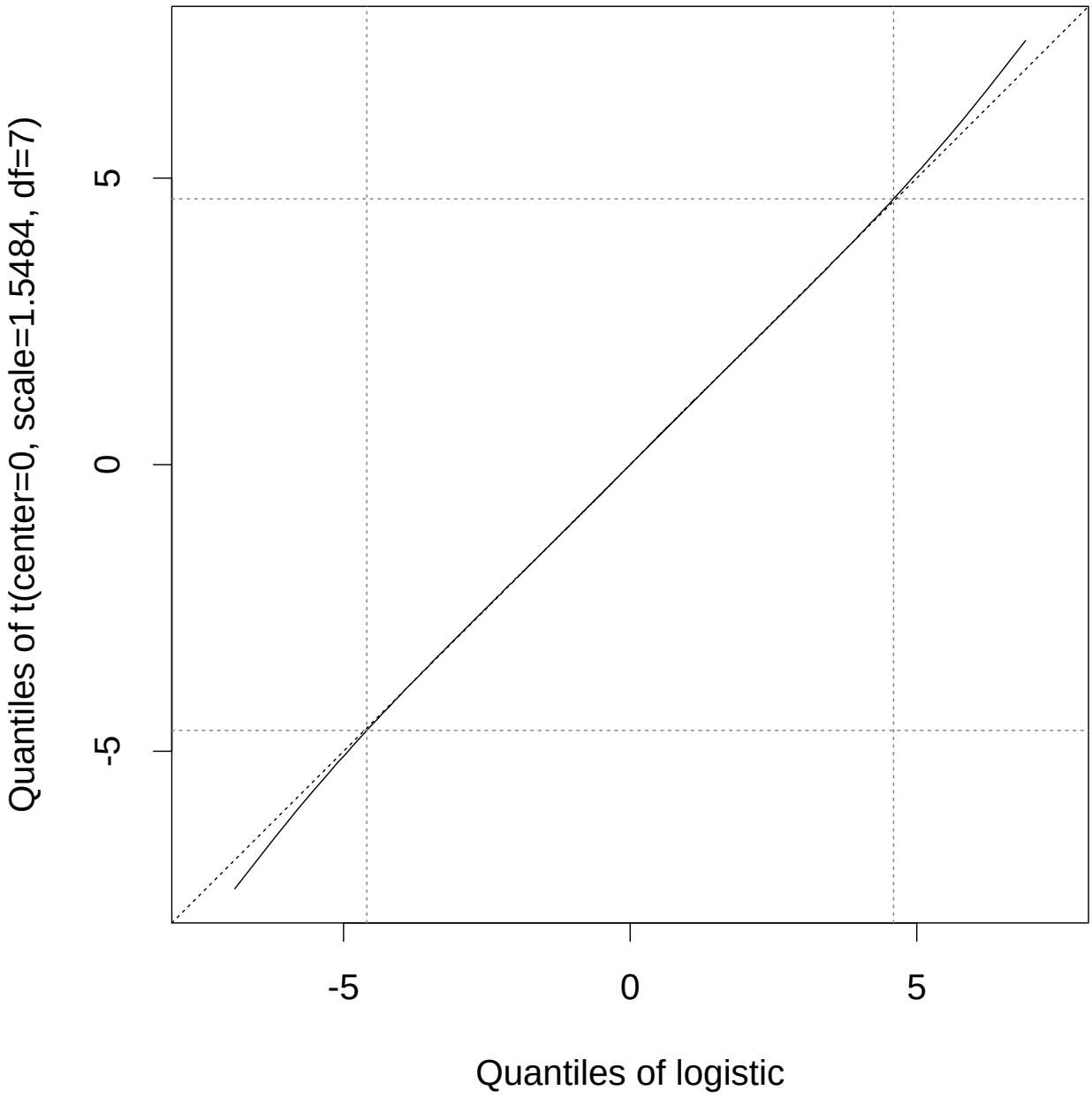

Figure 1: The Q-Q plot of the robit (7) model and the logistic model in the range corresponding to the probability range from 0.001 to 0.999. The horizontal and vertical dotted lines represent the 0.01 and 0.99 quantiles. The diagonal dotted line is the reference line indicating how well the two distributions match with each other.

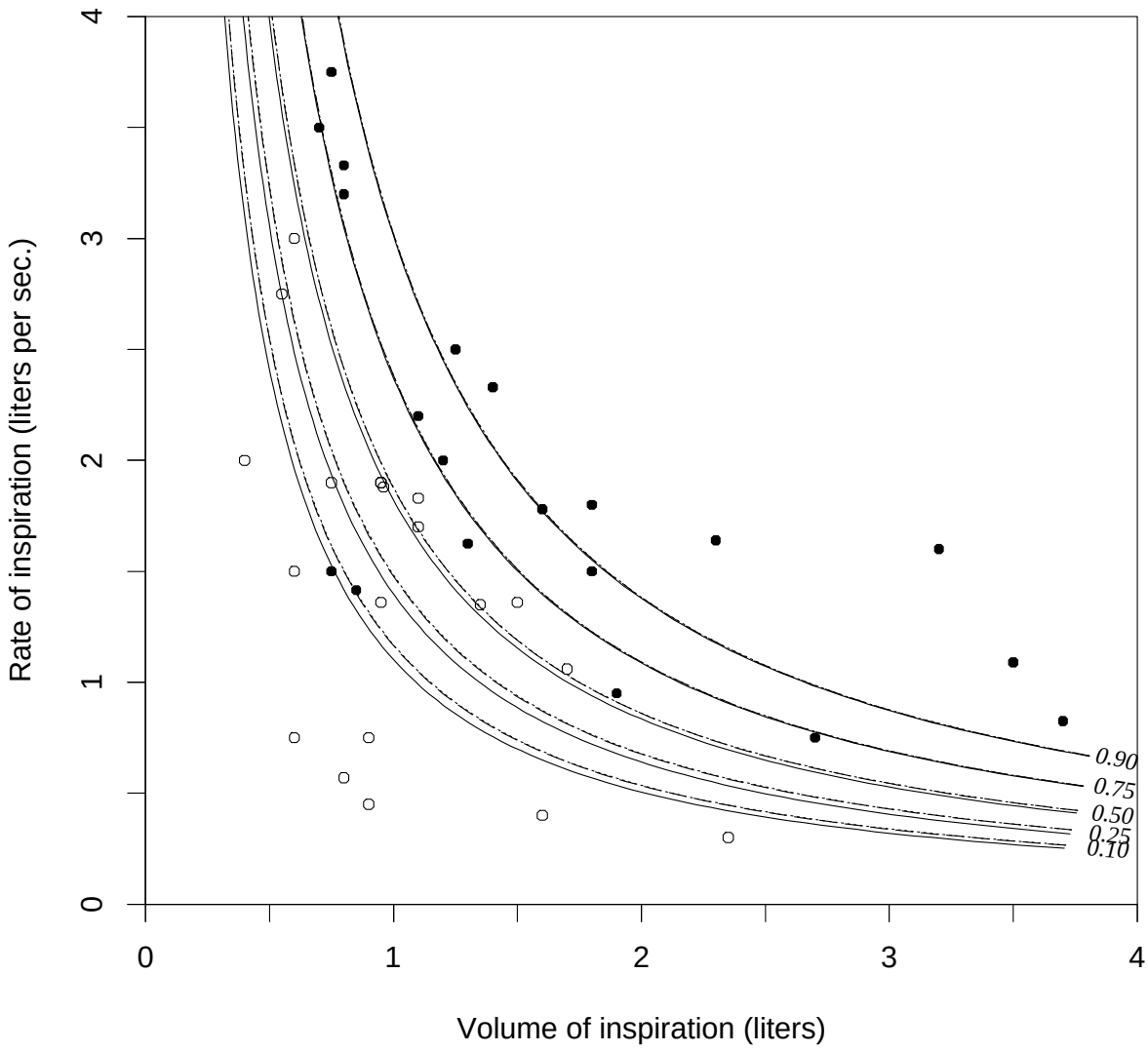

Figure 2: Scatter plot of the skin vaso-constriction data (with the symbols • and ◦ indicating positive and negative responses, respectively). The probability contours represent the probit (solid line), logistic (dotted line), and robit(7) (dashed line) models fitted by the methods of maximum likelihood.

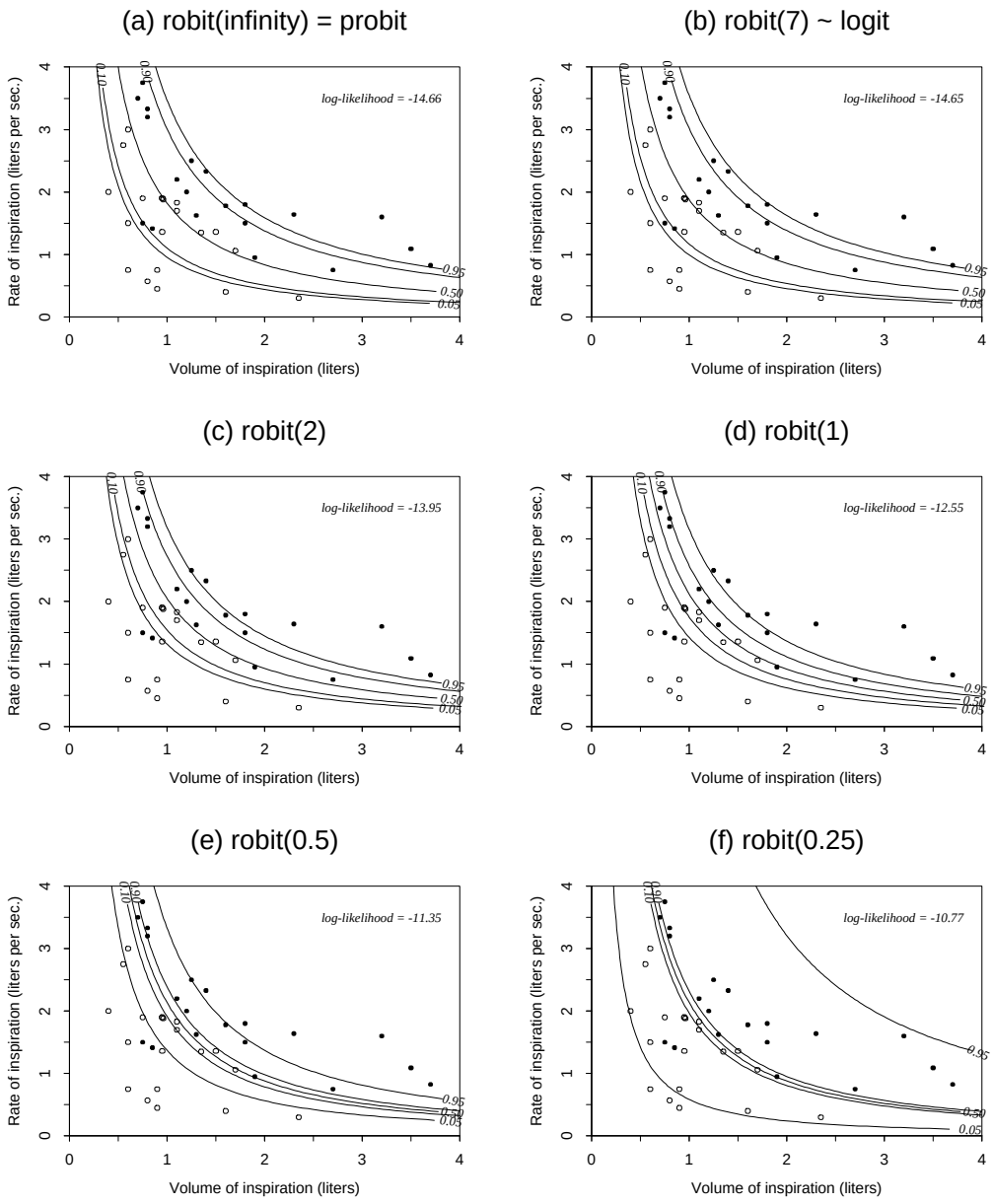

Figure 3: The robit models with various numbers of degrees of freedom fitted to the skin vasoconstriction data using the methods of maximum likelihood.

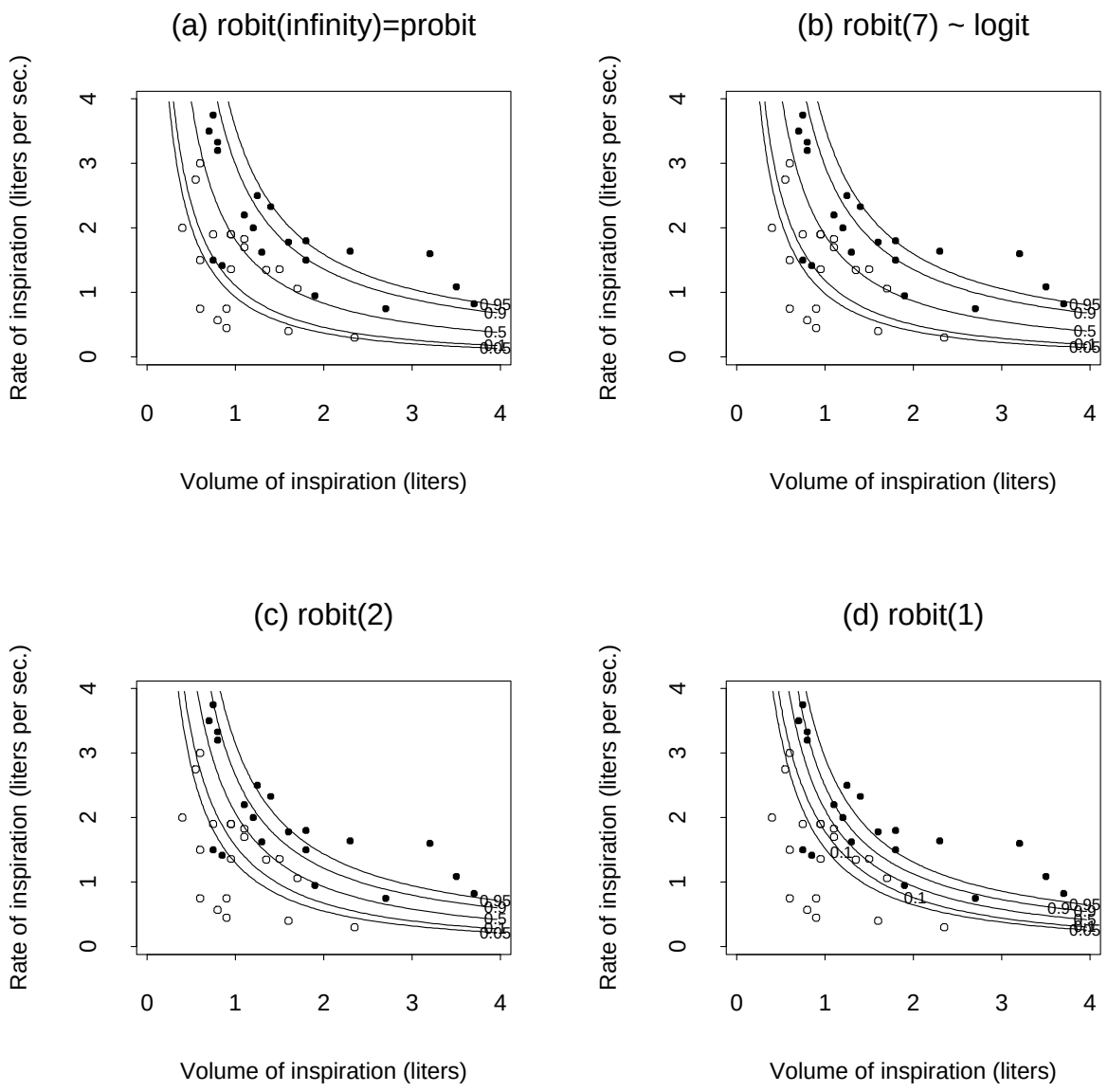

Figure 4: The robit models with various numbers of degrees of freedom fitted to the skin vasoconstriction data using the Bayesian methods.

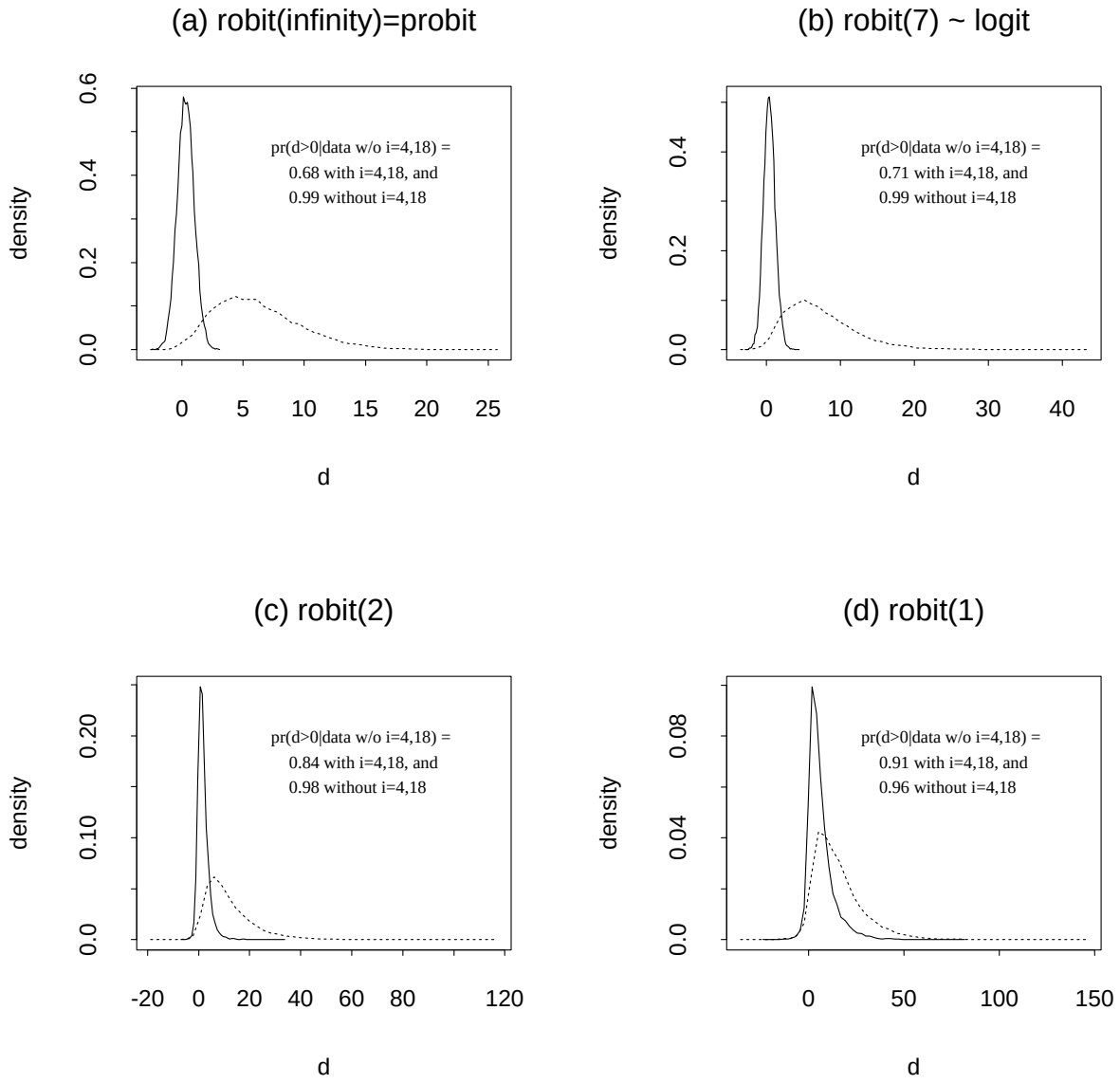

Figure 5: The posterior distributions of the difference $d = \beta_{\mathrm{RATE}} - \beta_{\mathrm{VOL}}$ obtained from the robit models with various numbers of degrees of freedom fitted to the skin vaso-constriction data with and without the two individual observations with $i = 4$ and 8.