A Note on Probit Model

Dawei Yin
Department of Computer Science & Engineering, Lehigh University
Bethlehem, PA 18015 USA
{day207}@cse.lehigh.edu

21 Sep. 2012

1 Probit Model

The pro bit model is used to solve binary response problem, that is, it can have only two possible outcomes which we will denote as 1 and 0.

$$Y = \begin{cases} 1 & \text{with probability } P \\ 0 & \text{with probability } 1 - P \end{cases}$$

The probit model rises when P is specific to be given by the normal cumulative distribution function evaluated at $X^T\beta$. That is $P = \Phi(X^T\beta)$.

$$P(Y = 1|X) = \Phi(X^T\beta)$$

where β are coefficient factors, which are also model parameters to estimate. The curves of the normal cumulative distribution function are shown in Fig **??**

It is also possible to motivate the probit model as a latent variable model which is often used to derive Gibbs sampling. Suppose there exist an auxiliary random variable

$$Y^* = X^T\beta + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, 1)$. Then Y can be viewed as an indicator for whether this latent variable is positive:

$$Y = \mathbf{1}_{\{Y^*>0\}} = \begin{cases} 1 \text{ if } Y^* > 0 \\ 0 \text{ otherwise} \end{cases}$$

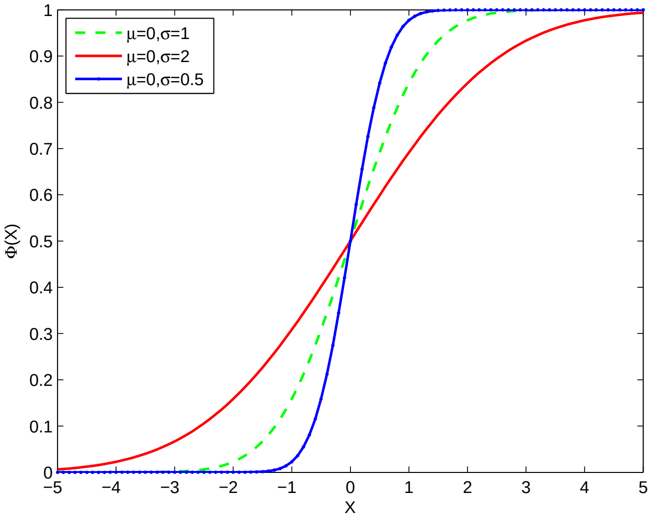


Figure 1: The curves of normal CDF

The use of the standard normal distribution causes no loss of generality compared with using an arbitrary mean and standard deviation because adding a fixed amount to the mean can be compensated by subtracting the same amount from the intercept, and multiplying the standard deviation by a fixed amount can be compensated by multiplying the weight by the same amount.

To see the two models are equivalent,

$$\begin{aligned} P(Y=1|X) &= P(Y^* > 0) = P(X^T\beta + \epsilon > 0) \\ &= P(\epsilon > -X^T\beta) \\ &= P(\epsilon < X^T\beta) \\ &= \Phi(X^T\beta) \end{aligned}$$

2 MLE for Probit model

For N observations $\{y_i, x_i\}_{i=1}^N$, the likelihood function is

$$L = \prod_{i_i}^{N} P_i^{y_i}(1-P_i)^{1-y_i}$$

where x_i is a feature vector for the ith observation. β are coefficient factors, which are also model parameters to estimate. Then the likelihood function can be rewritten as follows.

$$P(Y|X,\beta) = \prod_i^N \Phi(x_i^T\beta)^{y_i}(1-\Phi(x_i^T\beta))^{1-y_i}$$

The log-likelihood is

$$l = \ln L = \sum_i^N \left[y_i \ln \Phi(x_i^T\beta) + (1-y_i)\ln(1-\Phi(x_i^T\beta))\right]$$

Now, the first order conditions arising from Eqn. **??** are nonlinear and non-analytics. Therefore, we have to obtain the ML estimates using numerical optimization methods, e.g. gradient descent, the Newton-Raphson methods.

Note

$$\begin{aligned}
\phi(x) &= \frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2}) \\
\Phi(x) &= \int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})dx \\
\phi'(x) &= -x\phi(x) \\
\Phi(-x) &= 1-\Phi(x)
\end{aligned}$$

Then, we have

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^N \left[y_i\frac{\phi(x_i^T\beta)}{\Phi(x_i^T\beta)} - (1-y_i)\frac{\phi(x_i^T\beta)}{1-\Phi(x_i^T\beta)}\right]x_i$$

$$
\begin{aligned}
\frac{\partial^2 l}{\partial\beta\partial\beta'} &= \sum_{i=1}^{N}\left[y_i\frac{-x_i^T\beta\phi(x_i^T\beta)\Phi(x_i^T\beta)x_i-\phi^2(x_i^T\beta)x_i}{\Phi^2(x_i^T\beta)}\right. \\
&\left.-(1-y_i)\frac{-x_i^T\beta\phi(x_i^T\beta)(1-\Phi(x_i^T\beta))x_i+\phi^2(x_i^T\beta)x_i}{(1-\Phi(x_i^T\beta))^2}\right]x_i^T \\
&= -\sum_{i=1}^{N}\phi(x_i^T\beta)\left[y_i\frac{x_i^T\beta\Phi(x_i^T\beta)x_i+\phi(x_i^T\beta)x_i}{\Phi^2(x_i^T\beta)}\right. \\
&\left.+(1-y_i)\frac{-x_i^T\beta(1-\Phi(x_i^T\beta))x_i+\phi(x_i^T\beta)x_i}{(1-\Phi(x_i^T\beta))^2}\right]x_i^T \\
&= -\sum_{i=1}^{N}\phi(x_i^T\beta)\left[y_i\frac{x_i^T\beta\Phi(x_i^T\beta)+\phi(x_i^T\beta)}{\Phi^2(x_i^T\beta)}+(1-y_i)\frac{-x_i^T\beta(1-\Phi(x_i^T\beta))+\phi(x_i^T\beta)}{(1-\Phi(x_i^T\beta))^2}\right]x_ix_i^T
\end{aligned}
$$

It can be shown that this log-likelihood function is globally concave in ?, and therefore standard numerical algorithms for optimization will converge rapidly to the unique maximum.

Gradient Descent

$$\hat{\beta} \leftarrow \hat{\beta} + \eta\frac{\partial l}{\partial\beta}$$

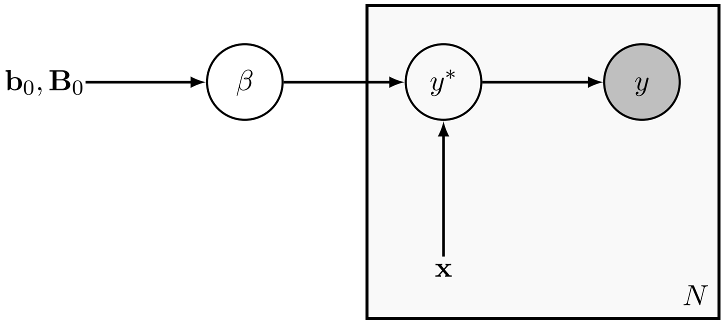
Newton's Method

$$\hat{\beta} \leftarrow \hat{\beta} - \left[\frac{\partial^2 l}{\partial\beta\partial\beta'}\right]^{-1}\frac{\partial l}{\partial\beta}$$

3 Gibbs Sampling for Probit model

Gibbs sampling of a probit model is possible because regression models typically use normal prior distributions over the weights, and this distribution is conjugate with the normal distribution of the errors (and hence of the latent variablesY*). The model can be described as

$$
\begin{aligned}
\beta &\sim \mathcal{N}(\mathbf{b}_0, \mathbf{B}_0) \\
y_i^*|\mathbf{x}_i, \beta &\sim \mathcal{N}(\mathbf{x}_i^T\beta, 1) \\
y_i &= \begin{cases} 1 \text{ if } y_i^* > 0 \\ 0 \text{ otherwise} \end{cases}
\end{aligned}
$$



For gibbs sampling, we have to calculate the conditional probability.

$$\begin{aligned} p(y^*|\mathbf{x}, \beta) &= \prod_{i=1}^{N} p(y_i^*|\mathbf{x}_i, \beta) \\ &= \prod_{i=1}^{N} \mathcal{N}(\mathbf{x}_i^T \beta, 1) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(y_i^* - \mathbf{x}_i^T \beta)^2\} \end{aligned}$$

$$p(\beta) = (2\pi)^{-\frac{k}{x}} |\mathbf{B}_0|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\beta - \mathbf{b}_0)^T \mathbf{B}_0^{-1} (\beta - \mathbf{b}_0)\}$$

$$
\begin{aligned}
p(\beta|y^*,\mathbf{x}) &\propto p(y^*|\mathbf{x},\beta)p(\beta) \\
&\propto |\mathbf{B}_0|^{-\frac{1}{2}}\exp\{-\frac{1}{2}\sum_{i=1}^{N}(y_i^*-\mathbf{x}_i^T\beta)^2-\frac{1}{2}(\beta-\mathbf{b}_0)^T\mathbf{B}_0^{-1}(\beta-\mathbf{b}_0)\} \\
&\propto |\mathbf{B}_0|^{-\frac{1}{2}}\exp\{-\frac{1}{2}\sum_{i=1}^{N}\left[y_i^{*2}-2y_i^*\mathbf{x}_i^T\beta+\mathbf{x}_i^T\beta\mathbf{x}_i^T\beta\right]-\frac{1}{2}(\beta-\mathbf{b}_0)^T\mathbf{B}_0^{-1}(\beta-\mathbf{b}_0)\} \\
&\propto |\mathbf{B}_0|^{-\frac{1}{2}}\exp\{\frac{1}{2}\sum_{i=1}^{N}2y_i^*\mathbf{x}_i^T\beta-\frac{1}{2}\sum_{i=1}^{N}\mathbf{x}_i^T\beta\mathbf{x}_i^T\beta-\frac{1}{2}(\beta-\mathbf{b}_0)^T\mathbf{B}_0^{-1}(\beta-\mathbf{b}_0)\} \\
&\propto |\mathbf{B}_0|^{-\frac{1}{2}}\exp\{\frac{1}{2}\left(\sum_{i=1}^{N}2y_i^*\mathbf{x}_i^T\beta-\sum_{i=1}^{N}\beta^T\mathbf{x}_i\mathbf{x}_i^T\beta-\beta^T\mathbf{B}_0^{-1}\beta+2\beta^T\mathbf{B}_0^{-1}\mathbf{b}_0\right)\} \\
&\propto |\mathbf{B}_0|^{-\frac{1}{2}}\exp\{\frac{1}{2}\left(\sum_{i=1}^{N}2y_i^*\mathbf{x}_i^T\beta-\beta^T\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T\beta-\beta^T\mathbf{B}_0^{-1}\beta+2\beta^T\mathbf{B}_0^{-1}\mathbf{b}_0\right)\} \\
&\propto |\mathbf{B}_0|^{-\frac{1}{2}}\exp\{\frac{1}{2}\left(\sum_{i=1}^{N}2y_i^*\mathbf{x}_i^T\beta-\beta^T(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T+\mathbf{B}_0^{-1})\beta+2\beta^T\mathbf{B}_0^{-1}\mathbf{b}_0\right)\} \\
&\propto |\mathbf{B}_0|^{-\frac{1}{2}}\exp\{\frac{1}{2}\left(\sum_{i=1}^{N}2y_i^*\mathbf{x}_i^T\beta-\beta^T\mathbf{B}^{-1}\beta+2\beta^T\mathbf{B}_0^{-1}\mathbf{b}_0\right)\} \\
&\propto |\mathbf{B}_0|^{-\frac{1}{2}}\exp\{\frac{1}{2}\left(2\beta^T(\sum_{i=1}^{N}y_i^*\mathbf{x}_i+\mathbf{B}_0^{-1}\mathbf{b}_0)-\beta^T\mathbf{B}^{-1}\beta\right)\} \\
&\propto |\mathbf{B}_0|^{-\frac{1}{2}}\exp\{\frac{1}{2}\left(2\beta^T\mathbf{B}^{-1}\mathbf{B}(\sum_{i=1}^{N}y_i^*\mathbf{x}_i+\mathbf{B}_0^{-1}\mathbf{b}_0)-\beta^T\mathbf{B}^{-1}\beta\right)\}
\end{aligned}
$$

$$\mathbf{B}=(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T+\mathbf{B}_0^{-1})^{-1}$$

Then the conditional probability of β is

$$\beta\sim\mathcal{N}(\mathbf{B}(\sum_{i=1}^{N}y_i^*\mathbf{x}_i+\mathbf{B}_0^{-1}\mathbf{b}_0),\mathbf{B})$$

$$\begin{aligned}
p(y|y^*, \beta, \mathbf{x}) &\propto p(y|y^*)p(y^*|\beta, \mathbf{x}) \\
&\propto \prod_{i=1}^{N} p(y_i|y_i^*)\mathcal{N}(y_i^*|\mathbf{x}_i^T\beta, 1) \\
&\propto \prod_{i=1,y_i=1}^{N} \mathbf{1}_{\{y_i^* \geq 0\}}\mathcal{N}(y_i^*|\mathbf{x}_i^T\beta, 1) \prod_{i=1,y_i=0}^{N} \mathbf{1}_{\{y_i^* < 0\}}\mathcal{N}(y_i^*|\mathbf{x}_i^T\beta, 1) \\
&\propto \prod_{i=1,y_i=1}^{N} \mathcal{N}^+(y_i^*|\mathbf{x}_i^T\beta, 1) \prod_{i=1,y_i=0}^{N} \mathcal{N}^-(y_i^*|\mathbf{x}_i^T\beta, 1)
\end{aligned}$$

Thus, the gibbs sampling

$$\begin{aligned}
\mathbf{B} &= (\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^T + \mathbf{B}_0^{-1})^{-1} \\
\beta|y^* &\sim \mathcal{N}(\mathbf{B}(\sum_{i=1}^{N} y_i^*\mathbf{x}_i + \mathbf{B}_0^{-1}\mathbf{b}_0) \\
y_i^*|y_i = 0, \mathbf{x}_i, \beta &\sim \mathcal{N}^-(y_i^*|\mathbf{x}_i^T\beta, 1) \\
y_i^*|y_i = 1, \mathbf{x}_i, \beta &\sim \mathcal{N}^+(y_i^*|\mathbf{x}_i^T\beta, 1)
\end{aligned}$$

Acknowledgments

References