PROBIT REGRESSION

FRANK WOOD

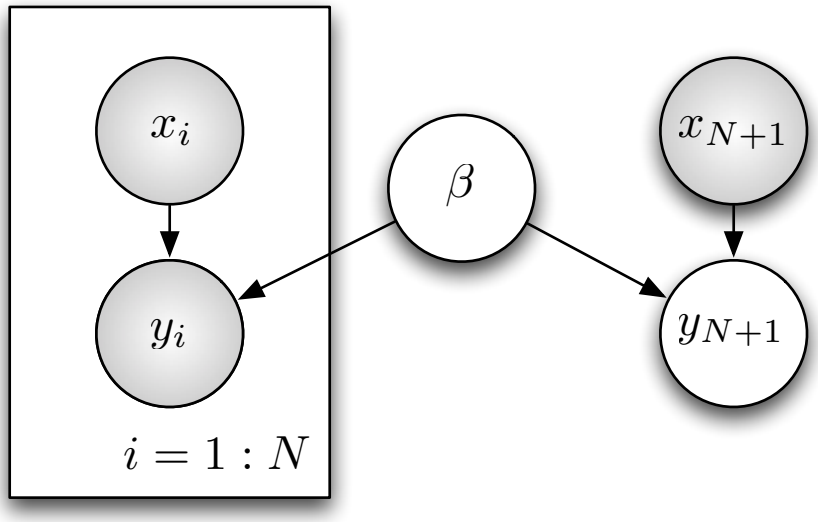

FIGURE 1. Probit model, no auxiliary variables

For reasons that are somewhat obscure to me, statisticians tend to use probit regression for binary classification whereas machine learners tend to use logistic regression. In a recent paper, my collaborators and I found it useful for computational purposes to use probit regression. One can find many good primers on probit regression around the web, but, as we all know, there is almost always space for another.

Figure 2 is the cumulative distribution function (cdf) of a $N(0, 1)$ distribution. The "probit" function is the inverse of the normal cdf. The normal cdf function, denoted $\Phi(x; \mu, \sigma^2)$ with μ the mean, σ^2 the variance and x the argument, is actually much more relevant in this context.

The range of the normal cdf is $(0, 1)$ which means that it can be interpreted as a probability. For instance, one can construct a generalized linear model (a "probit regression model") of the form

$$P(y_i = 1) = \Phi(x_i^T \beta; 0, \sigma^2). \tag{1}$$

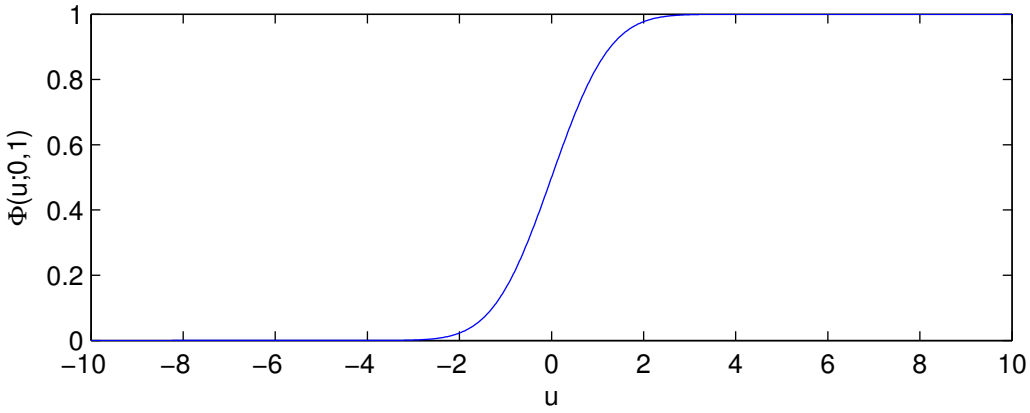

FIGURE 2. CDF of $N(0,1)$

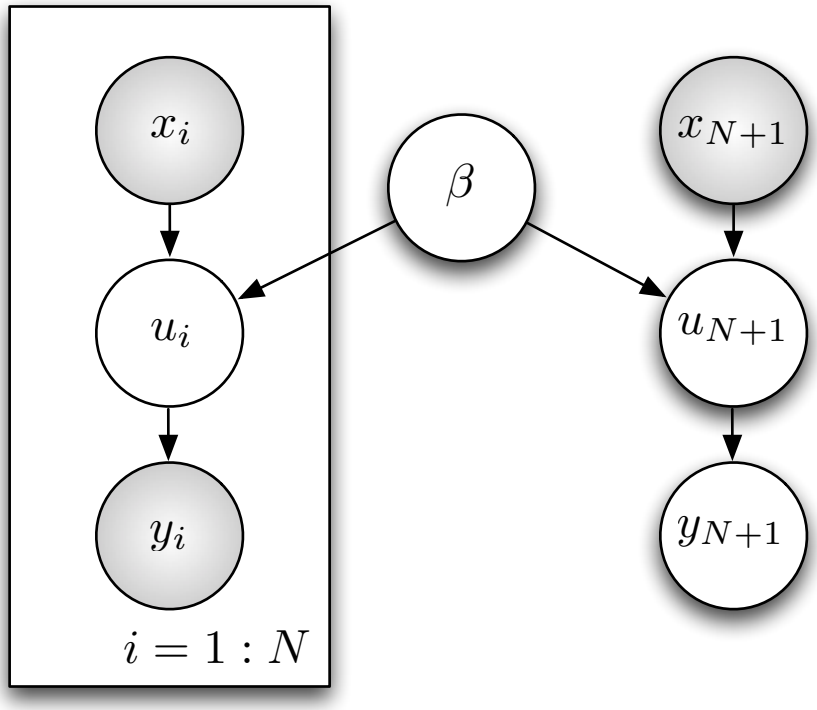

FIGURE 3. Probit model

Depending on convention (i.e. binary y_i represented as $\{1, 0\}$ or $\{1, -1\}$) then the probability of y_i being labeled the opposite way is $P(y_i = -1)$ or $P(y_i = 0) = 1 - P(y_i = 1)$. Here x_i is a vector of covariates, β is a vector of weights, and y_i is a single, binary valued response. As always, the close relationship between regression and classification are in full display here : probit regression is a "generalized linear *regression* model" and a "binary classifier."

Figure 1 shows the graphical model for probit regression (minus any priors on the regression weights β). In this model we would like to use labeled training data, $\{x_i, y_i\}_{i=1}^N$ to "learn" the value of β and then to use this value to predict the value of $y_{N+1}|x_{N+1}, \beta$. We will be being Bayesian about our inference so here, what we really mean is, we will average over the posterior distribution of β when making predictions. This means that we want to draw samples from the posterior distribution of $\beta|\{x_i, y_i\}_{i=1}^N$. This brings us to Figure 3 which introduces a set of auxiliary variables $\{u_i\}_{i=1}^N$. In this note we will demonstrate that the model in Figure 3 is the same as the model in Figure 1 when the u_i's are marginalized out and will suggest that inference in the former is computationally easier.

First – what is an auxiliary variable? It is a variable introduced into a model in order to make inference easier but whose existence does not change the distribution of interest. Auxiliary variables for slice sampling are one particularly clever use of auxiliary variables. The auxiliary variable trick in probit regression is another.

For the purposes of exposition, let's forget about the i index for a second and focus on a single instance y, x, and u. The argument we make will hold for all by simply reintroducing subscripts.

To start, let's write down the joint distribution of these quantities (according to the graphical model that includes auxiliary variables).

$$P(y, x, u) = P(y|u)P(u|x, \beta) \tag{2}$$

Clearly, straight away, one can see precisely why this auxiliary variable scheme works. By the law of total probability we have

$$P(y, x) = \int P(y, x, u)du = \int P(y|u)P(u|x, \beta)du. \tag{3}$$

If by MCMC we generate S samples $\{u^{(s)}, y^{(s)}, x^{(s)}\}_{s=1}^S \sim P(y, x, u)$ we know that marginalizing u out (i.e. disregarding its value) we get samples $\{y^{(s)}, x^{(s)}\}_{s=1}^S \sim P(y, x)$.

We haven't specified the most important part of the auxiliary variable sampling scheme yet, namely, what $P(y|u)$ and what $P(u|x, \beta)$ are. Let's try $y = \text{sign}(u)$ and $u \sim N(x^T\beta, \sigma^2)$. These choices are nice in a particular way. First let's verify that the marginalization of u out of this model results in the model specification in Equation 1.

$$\begin{aligned} P(y=1|x,\beta) &= \int P(y=1|u)P(u|x,\beta)du \\ &= \int \mathbb{I}(u>0)N(u;x^T\beta,\sigma^2)du \\ &= \int_0^\infty N(u;x^T\beta,\sigma^2)du \\ &= 1-\Phi(0;x^T\beta,\sigma^2) \\ &= \Phi(x^T\beta;0,\sigma^2) \end{aligned}$$

where the last line comes from the fact that for symmetric distributions like the normal distribution, $\Phi(x^T\beta;0,\sigma^2) = 1-\Phi(-x^T\beta;0,\sigma^2)$ and the mean of a normal cdf can be translated arbitrarily, i.e. $\Phi(-x^T\beta;0,\sigma^2) = \Phi(0;x^T\beta,\sigma^2)$ (which comes from adding the offset $x^T\beta$ to the cdf argument and mean).

OK. So, now, we have established the fact that for a particular sort of auxiliary variable choice, we get the same probit model as we wanted. Why is this choice nice?

Well, it comes down to sampling β and u and y. Generally, sampling β in the model without auxiliary variables will require hybrid Monte Carlo (HMC) or Metropolis Hastings of some sort. Gibbs sampling often comes with substantial benefits. By making this choice of auxiliary variable the conditional distribution of u_i given everything else is proportional to a truncated Normal distribution, a distribution that is, by nature of its commonness, relatively straightforward to sample from. The big benefit, though, acrues from the posterior form for sampling β. With the u's "observed" (as they would be in a Gibbs sampler), the posterior distribution of β (for typical choices of prior) is precisely the same as that for linear regression, perhaps the most well studied model in statistics. Sampling β from its posterior distribution typically is quite simple; certainly more so that sampling β without the u auxiliary variables.