

# A Note On Fitting the Probit Model in SAS

G.J. Johnston, SAS Institute Inc., Cary, NC

March 16, 1993

This is an informal description of the method used to fit a probit model in the SAS PROBIT procedure.

The SAS PROBIT procedure uses a modified Newton-Raphson algorithm to fit the probit model. Suppose, for simplicity, that the data are binary (success-failure) and that there is a single covariate  $x$ . Then the probit model for the probability of success is

$$p_i = \Phi(a + bx_i)$$

for observation  $i$  where  $a$  and  $b$  are unknown parameters to be estimated and  $\Phi$  is the standard normal cumulative distribution function. PROC PROBIT fits more complicated models (multinomial response, threshold parameter, and more covariates), but this simpler model extends directly to more complicated ones and should give you an idea of how SAS fits the probit model.

The Newton-Raphson step for updating the parameter vector  $\theta^{(j)}$  at the  $j$ th step in the iterative fitting process is

$$\theta^{(j+1)} = \theta^{(j)} - \mathbf{H}^{-1} \mathbf{g}$$

where  $\mathbf{H}$  is the matrix of second derivatives of the log likelihood function with respect to the parameters and  $\mathbf{g}$  is the vector of first derivatives. The likelihood function is

$$L = \prod_i \binom{n_i}{r_i} p_i^{r_i} (1 - p_i)^{n_i - r_i}$$

and after dropping the binomial coefficient which has no effect on parameter estimates or covariances, the log likelihood is

$$l = \sum_i [r_i \log(p_i) + (n_i - r_i) \log(1 - p_i)]$$

where  $n_i$  is the number of trials and  $r_i$  is the number of successes on observation  $i$ , and the product and sum are over observations.

You need to compute the first and second derivatives of the log likelihood function to implement the Newton-Raphson algorithm. For any parameter  $\theta$ , the derivatives are

$$\frac{\partial l}{\partial \theta} = \sum_i \left[ \frac{r_i}{p_i} - \frac{n_i - r_i}{1 - p_i} \right] \frac{\partial p_i}{\partial \theta}$$

and

$$\frac{\partial^2 l}{\partial \theta_1 \partial \theta_2} = \sum_i \left[ - \left( \frac{r_i}{p_i^2} + \frac{n_i - r_i}{(1 - p_i)^2} \right) \frac{\partial p_i}{\partial \theta_1} \frac{\partial p_i}{\partial \theta_2} + \left( \frac{r_i}{p_i} - \frac{n_i - r_i}{1 - p_i} \right) \frac{\partial^2 p_i}{\partial \theta_1 \partial \theta_2} \right]$$

Note that, from the definition of the probit model,

$$\frac{\partial p_i}{\partial a} = \phi(a + bx_i)$$

$$\frac{\partial p_i}{\partial b} = x_i \phi(a + bx_i)$$

$$\frac{\partial^2 p_i}{\partial^2 a} = -(a + bx) \phi(a + bx)$$

and so on, where  $\phi$  is the standard normal probability density function. Inserting these into the expressions for the derivatives of the log likelihood function and replacing  $a$  and  $b$  with their current estimates gives the formulas to use in the Newton-Raphson step. This is the basic iteration step of the algorithm that SAS uses to fit the probit model.

SAS estimates the covariance matrix of the parameter estimates with the inverse of the negative of the matrix of second derivatives, or the inverse of the *observed* information matrix, evaluated at the last iteration.

Notice that the expected value of the second derivative is

$$E \frac{\partial^2 l}{\partial \theta_1 \partial \theta_2} = \sum_i \left[ \frac{n_i}{p_i(1 - p_i)} \frac{\partial p_i}{\partial \theta_1} \frac{\partial p_i}{\partial \theta_2} \right]$$

You can replace the matrix of second derivatives with the matrix of expected values of second derivatives and use this in the Newton-Raphson algorithm. The covariance matrix is then estimated by the inverse of the *expected* information matrix. Finney uses this method, and some statistical computing packages also use it.

The two methods will generally converge to the same parameter estimates, but the estimates of the covariance matrix will be slightly different.