Statistics Project

Federico Marinozzi
Leonardo Polverari
Mattia Angius

January 2024

1 Probit regression

Probit regression is a generalized linear model characterized by the following components:

- random component: $Y_i \sim Bernoulli(\mu_i)$, with i = 1, ..., n (number of obs)
- systematic component: $\eta_i = X_i^T \beta$
- link function: $\Phi^{-1}(\mu_i) = \eta_i$

Since it can be shown that the Bernoulli distribution belongs to the exponential family:

$$f(y_i\,;\theta_i,\phi) = e^{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)}$$

as

$$f(y_i\,;p_i) = p_i^{y_i}(1-p_i)^{1-y_i} = e^{y_i \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i)},$$

with

$$\theta_i = \log\left(\frac{p_i}{1-p_i}\right), \quad b(\theta_i) = \log\left(1+e^{\theta_i}\right), \quad a(\phi) = 1, \quad c(y,\phi) = 0,$$

it follows that

$$\mu_i = E[Y_i] = p_i = \frac{e^{\theta_i}}{1+e^{\theta_i}}, \quad \mathrm{Var}(Y_i) = p_i(1-p_i) = \mu_i(1-\mu_i) = \frac{e^{\theta_i}}{(1+e^{\theta_i})^2}.$$

Moreover, since the link function is given by $\Phi^{-1}(\mu_i) = \eta_i$, we have that

$$\mu_i = \Phi(\eta_i).$$

The last relevant quantity to explicit is

$$\frac{\partial \eta_i}{\partial \mu_i} = \Phi^{'}(\mu_i) = \frac{1}{\Phi'(\Phi^{-1}(\mu_i))} = \frac{1}{\phi(\eta_i)}.$$

where ϕ is the pdf of the normal distribution. The likelihood function of the Bernoulli is:

$$L(\beta\,|\,Y,X) = \prod_{i=1}^{n} \left[\mu_i^{y_i}(1-\mu_i)^{1-y_i}\right]$$

that expressed in terms of η becomes:

$$L(\beta\,|\,Y,X) = \prod_{i=1}^{n} \left[\Phi(\eta)^{y_i}\left(1-\Phi(\eta)\right)^{1-y_i}\right]$$

2 Fisher Scoring method

2.1 Algorithm

The Newton's method is an optimization algorithm that consists of the following steps:

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}$$

where $f(x_n)$ is the function we want to minimize, $f'(x_n)$ is the first derivative (gradient) of the function with respect to x and $f''(x_n)$ is the second derivative (Hessian).

1. Choose an initial guess x_0.
2. For each iteration n, update x_{n+1} using the equation above
3. Repeat until convergence: $|x_{n+1} - x_n| < \epsilon$, where ϵ is a small tolerance.

The Fisher Scoring method is a statistical application of the Newton's method, where the function to minimize is the "minus" log-likelihood function and the updating equation becomes:

$$\theta_{t+1} = \theta_t + I_n(\theta_t)^{-1}\nabla l(\theta_t),$$

where θ is the set of parameters to estimate, I_n is the expected Fisher information and l is the log likelihood function.

2.2 Application to Probit

Calling b_t the set of parameters we want to estimate, it can be shown that the updating equation can be rewritten as:

$$b_{t+1} = \left(X^t W_t X\right)^{-1} X^t W_t Z_t,$$

where W is a $n \times n$ diagonal matrix with elements

$$W_{ii} = \mathrm{Var}(Y_i)^{-1} \left(\frac{\partial \eta_i}{\partial \mu_i}\right)^{-2}$$

That in the case of the Probit model, from what has been shown in section 1, become:

$$W_{ii} = \frac{\phi(\eta_i)^2}{\mu_i(1-\mu_i)}$$

and Z is a $n \times 1$ vector with elements (n = number of obs)

$$Z_i = \eta_i + (Y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)$$

That in the case of Probit become:

$$Z_i = \eta_i + \frac{Y_i - \mu_i}{\phi(\eta_i)}.$$

2.3 Convergence

The algorithm converges when the estimated values of b stabilize around a certain value. When fitting the SAheart data we noticed that convergence was reached early with the log-likelihood stabilizing after only two iterations around the value of -236 (*Fig1*) and the parameters stabilizing after three iterations (*Fig2*).

3 Random Walk Metropolis Hastings

3.1 Algorithm

The Random Walk Metropolis Hastings is an algorithm used to generate a series of samples from a target probability distribution, incorporating prior beliefs regarding the distribution of the parameter to estimate and using a transition kernel (a proposal distribution).
With β being the set of parameters we want to estimate, $\pi(\beta)$ being the prior distribution of β, $\pi(\beta|y) = \frac{\pi(\beta)L(\beta,y)}{f(y)}$ (where L is the likelihood function) being the posterior distribution and $g(\beta|\beta_{t-1})$ being the transition kernel, the algorithm does the following:

1. choose a starting point β_0
2. draw $\beta^\star$ from $g(\beta|\beta_{t-1})$
3. evaluate the probability ratio $\alpha = min\{1, \frac{\pi(\beta_\star|y)g(\beta_\star|\beta_t)}{\pi(\beta_t|y)g(\beta_t|\beta_\star)}\}$
4. draw u from a Uniform distribution $u \sim U(0,1)$
5. accept $\beta_\star$ as β_{t+1} if $\alpha >= u$, else $\beta_{t+1} = \beta_t$
6. add β_t to the chain
7. repeat the steps from 2 to 6 for the desired number of iterations

After the chain is built two more steps can be performed:

1. burn-in: discard an initial portion of the chain in order to make the chain converge to the target distribution in an easier way.
2. thinning: in order to reduce autocorrelation, retain only every k-th sample.

If the obtained chain $\{\beta_t\}$ is irreducible and aperiodic, it holds that:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \beta_t = E[\beta|\mathrm{y}]$$

where $E[\beta|\mathrm{y}]$ represents the expected value of the parameter conditional on the observed data.

3.2 Application to Probit

3.2.1 Prior distribution

The choice of a prior distribution is a crucial step in the implementation of the Metropolis Hastings algorithm. In a Bayesian setting, probability is used as a tool to formalize our uncertainty on a phenomenon. As a fundamental aspect of the learning process, it is therefore paramount to describe our initial beliefs on the distribution of the parameter of interest. Given a target parameter θ , its prior distribution can be formally defined as:

$$\theta \sim \pi(\cdot) \quad : \quad \int_{\Theta} \pi(\theta)\, d\theta = 1$$

The decision between an uninformative and an informative prior will have a significant impact on the posterior distribution. Furthermore, it will affect the complexity of the algorithm, as well as the time of execution. To maximize the computational efficiency we considered two priors: a (multivariate) normal one with a mean of 0 (null vector) and a variance of 1 (identity matrix), and an uninformative prior distribution. The former allows us to express uncertainty without imposing strong assumptions, exploiting our beliefs regarding the small magnitude of the coefficients. The latter strengthens the notion of our initial lack of beliefs, by assuming that each parameter is equally likely. Given the negligible differences in coefficients between the two priors, we will present and discuss our results using the multivariate normal one as our point of reference.

3.2.2 Transition kernel

The considerable number of parameters to estimate and their small magnitude led the choice of the transition kernel towards a multivariate normal centered in β_t, with $\Sigma = \tau I$, where $\tau = 0.0006$. Such a small choice of τ can be explained by the small dimension of the coefficients and the consequent risk of obtaining a too small likelihood function after few iterations.

3.2.3 Probability ratio

Given these choices, the probability ratio becomes:

$$\alpha = min\{1, \frac{\pi(\beta_\star|y)}{\pi(\beta_t|y)}\}$$

Since g is simmetric and therefore $g(\beta_t|\beta_\star) = g(\beta_\star|\beta_t)$. Given the Probit likelihood and the prior distribution, α becomes:

$$\alpha = min\{1, \frac{\pi(\beta_\star)L(\beta_\star, y)}{\pi(\beta_t)L(\beta_t, y)}\} = min\{1, \frac{\prod_{i=1}^{n}\left[\Phi(\eta_\star)^{y_i}\left(1-\Phi(\eta_\star)\right)^{1-y_i}\right] pdf(\beta_\star)}{\prod_{i=1}^{n}\left[\Phi(\eta_t)^{y_i}\left(1-\Phi(\eta_t)\right)^{1-y_i}\right] pdf(\beta_t)}\}$$

3.3 Convergence

While the convergence of the Fisher scoring algorithm was straightforward, the Random Walk Metropolis Hastings had some problems. Indeed, while fitting the data convergence was reached only for the distributions of some coefficients, with other chains being clearly non-stationary. Even changing prior distribution and iterating a high number of times with different τ. For this reason, we tried to use our model to fit simulated data and obtained great results both in terms of convergence and estimation. We then thought that because of the fact that the Probit model is based on the inverse normal CDF, scaling the data could have helped tackle the issue. The scaling produced indeed great results in terms of stationarity of the chains, however, it probably hinders the interpretability of the coefficients. For this reason, we still decided to perform our coefficient analysis considering the ones obtained fitting the unscaled data using 500.000 iterations with $\tau = 0.0006$. As *Fig 3* shows, we obtained relatively stationary chains keeping the last 150.000 observations (save for the constant term, whose magnitude is probably much higher than the extimated one) and performing thinning three times (even though some acf follow a weird path).

4 Coefficient interpretation

Interpreting the coefficients in a simple linear regression is a straightforward task: given an outcome variable and a set of regressors, a specific coefficient represents the change in the dependent variable as a consequence

of a one unit increase in the dependent variable, keeping everything else equal. In the case of binary models, however, the state of interest we are analyzing is the result of an unobservable model surpassing a certain threshold. Therefore, the data under analysis provide information on the outcome of interest only through their effect on the latent model. Furthermore, the marginal effect of one regressor is dependent on a function f of all other explanatory variables. Formally:

$$\frac{\delta \Pr(y_i = 1|x_i)}{\delta x_{i,j}} = \frac{\delta F(x_i'\beta)}{\delta x_i'\beta \cdot \beta_j} = \beta_j f(x_i'\beta)$$

Where f is the derivative of the cdf of the standard normal.

The dataset we analyzed comprises 462 observations on variables that might cause the insurgence of coronary heart diseases. The 10 regressors are the following:

- Systolic blood pressure (sbp)
- Cumulative use of tobacco, in kg
- Low density lipoprotein cholesterol (LDL)
- Levels of adiposity, in the form of a numeric vector
- A binary variable indicating the presence or absence of a family history of heart disease (FamHist)
- A scale indicating the degree of strength of a Type A behavior. High ratings correspond to increased competitiveness, ambitiousness, stress, impatience etc..
- Obesity
- Alcohol consumption
- Age

After properly cleaning, rearranging and fitting the data, we ran the Fisher Scoring and the Random Walk Metropolis Hastings algorithm to identify the coefficients of interest.

Algorithm	FS	MH
Intercept	-3.570	-0.162
Sbp	0.003	-0.005
Tobacco	0.048	0.056
LDL	0.103	0.100
Adiposity	0.012	0.033
FamHist	0.539	0.605
TypeA	0.023	0.005
Obesity	-0.040	-0.091
Alcohol	1.96e-05	-4e-06
Age	0.026	0.018

Table 1: Coefficients results with Fisher Scoring and Metropolis Hastings algorithms

As indicated by the results, the differences between the coefficients obtained with the two algorithms are negligible. Furthermore, our results solely provide point estimations. For certain predictors, such as Alcohol or Sbp, it could therefore be argued that the coefficients may not significantly differ from 0.

In both the models, it is evident that all predictors, except for obesity, have a positive influence on the likelihood of developing a coronary heart disease. The regressor with the most significant coefficient magnitude seems to be the presence of a family history of coronary heart diseases. Following closely are low-density lipoprotein cholesterol and tobacco consumption, respectively.

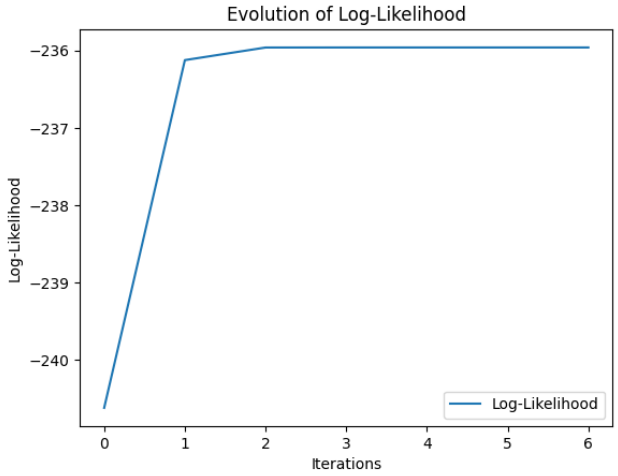

Figure 1: Log Likelihood convergence in Fisher Scoring

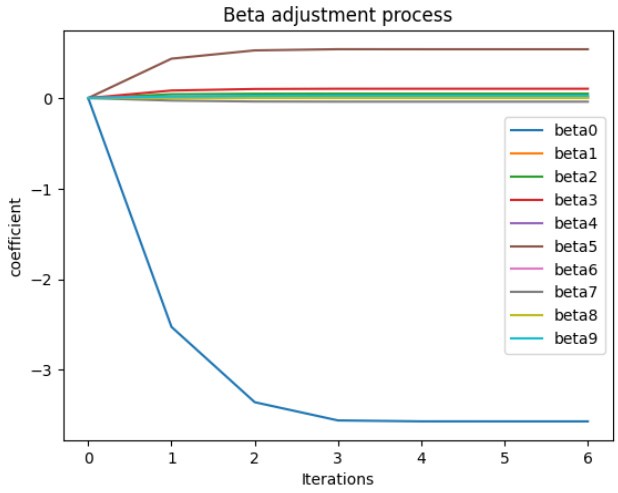

Figure 2: Convergence of coefficients in Fisher Scoring

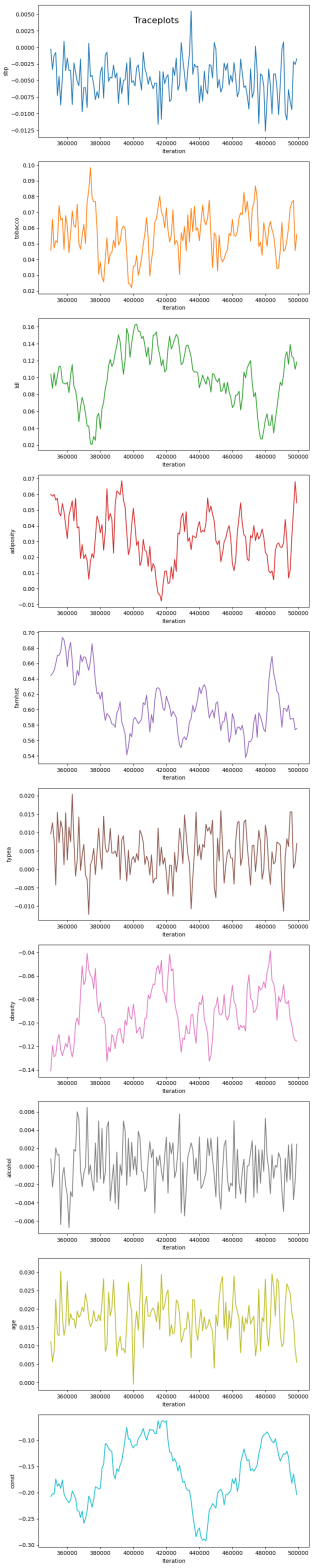

Figure 3: Coefficients in RWMH after burn-in and thinnin