Computational Statistics
Final Project

20592 – Statistics and Probability

Data Science and Business Analytics
2023/2024

Made by:
Valentina Brivio
Pasquale Caponio
Enrico Cipolla Cipolla
Mariapia Tedesco
Arianna Zottoli

1. Generalized Linear Models and Probit Regression

Let us a binomial Generalized Linear Model, which components are:

- Random component: $f(y_i, \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c\right\}$
- Systematic component: $\eta_t = \beta_0 + \sum_{i=1}^{p} \beta_j x_{ij} = X^T\beta$
- Link function: $g(\mu_i) = \eta_i, \quad \mu_i = E[y_i|x_i]$

The linear component $\eta_i = x_i^T\beta$ is linked to the expected value μ_i through a link function $g(\cdot)$: $g(\mu_i) = x_i^T\beta$. When dealing with binary response data, the random component takes the form of a Bernoulli distribution and we talk about Binomial GLMs. For binomial GLMs, specific link functions are required to ensure that $0 < \mu < 1$. Therefore, a monotonically increasing function $g: [0,1] \to \mathbb{R}$ is chosen.

In this case the of Probit regression, the link function is given by the inverse of the cumulative distribution function of a Gaussian random variable: $g(\mu_i) = \Phi^{-1}(\mu_i)$, from which we get $\mu_i = \Phi(x_i^T\beta)$.
Our goal is to find the vector of parameters β that maximizes the log-likelihood function $l(\beta, x)$, which means we want to find $\underset{\theta}{\operatorname{argmin}} -l(\beta, x)$.

1.1 Fisher Scoring Algorithm

The Fisher Scoring algorithm, employed in maximum likelihood estimation, iteratively refines parameter estimates. It begins with an initial guess for the parameters to be estimated, which in our algorithm is set to 0. Then, at each iteration the algorithm computes the following parameters:

$$\eta_t = X^T\beta_t, \qquad \mu_t = g^{-1}(\eta_t), \qquad Z_t = \eta_t + (y - \mu_t)\left(\frac{\partial\eta_t}{\partial\mu_t}\right), \qquad W_t = \frac{1}{V(y)}\left(\frac{\partial\eta_t}{\partial\mu_t}\right)^{-2}$$

And lastly:

$$\beta_{t+1} = (X^T W_t X)^{-1} X^T W_t Z_t$$

The algorithm updates iteratively the parameters, ultimately converging to the maximum likelihood estimates. A convergence criterion is therefore included in the algorithm and is checked at each iteration. It may be based on the absolute or relative change in parameter values, or it might involve monitoring the log-likelihood function's behavior. In the main case of our algorithm (absolute and relative convergence can be found in the Jupiter notebook), we implemented a modified relative convergence criterion which computes the distance between the old and new estimate divided by the sum between the old estimate and a predetermined epsilon ε (set to 0.000001). Then, if such computation is below the epsilon ε threshold, convergence is achieved. Once the convergence criterion is met, the algorithm terminates, and the final estimates the parameters β represent the maximum likelihood solution.

1.2 Analytical Derivation

As previously states, our goal is in fact to find $\underset{\theta}{\operatorname{argmin}} -l(\beta, x)$.
To do so we can use an iterative procedure of optimization, among which the Newton's method:

$$b_{t+1} = b_t - \frac{l\prime(b_t)}{l''(b_t)}.$$

We can substitute $-l''(b_t)$ with the observed Fisher Information and the equation becomes

$$b_{t+1} = b_t - \frac{l\prime(b_t)}{\mathbb{I}(b_t)}.$$

After some computation, we deduce that

$$b_{t+1} = (X^T W_t X)^{-1} X^T W_t Z_t$$

where:

- W is the diagonal matrix $(n \times n)$ with elements $W_{ii} = Var(Y_i)^{-1}\left(\frac{\delta\eta_i}{\delta\mu_i}\right)^{-2}$;
- Z is a vector $(n \times 1)$ with elements $Z_i = \eta_i + (Y_i - \mu_i)\left(\frac{\delta\eta_i}{\delta\mu_i}\right)$;
- X is the covariates matrix $(n \times p)$.

Since this is the case of binomial GLM: $Var(Y_i) = \mu_i(1 - \mu_i)$;

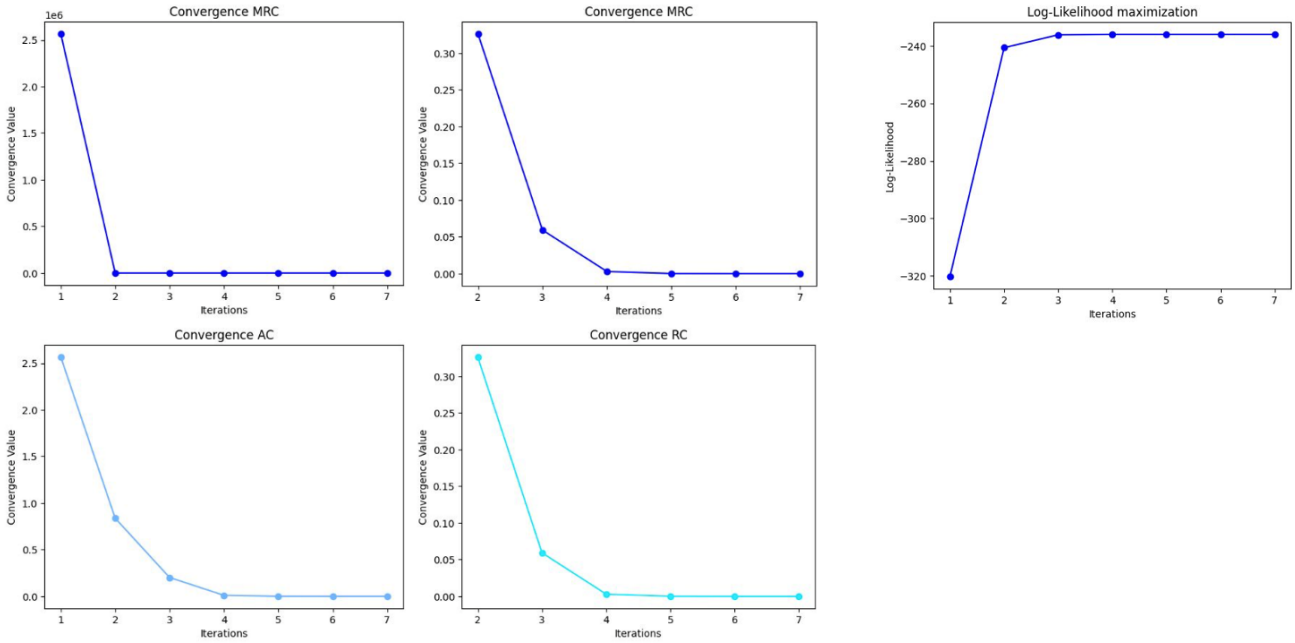
and since we have chosen a probit regression: $\mu_i = \Phi(x_i^T \beta)$ and $\frac{\delta \eta_i}{\delta \mu_i} = \frac{\delta \Phi^{-1}(\mu_i)}{\delta \mu_i} = \frac{1}{\phi(\Phi^{-1}(\mu_i))}$.
We can conclude that:

$$W_{ii} = Var(Y_i)^{-1} \left(\frac{\delta \eta_i}{\delta \mu_i}\right)^{-2} = \frac{\left[\phi\left(\Phi^{-1}(\mu_i)\right)\right]^2}{\mu_i(1-\mu_i)};$$
$$Z_i = \eta_i + (Y_i - \mu_i)\left(\frac{\delta \eta_i}{\delta \mu_i}\right) = x_i^T \beta + (Y_i - \mu_i)\left[\phi\left(\Phi^{-1}(\mu_i)\right)\right]^{-1}.$$

1.3 Convergence diagnostics

Convergence ensures that the algorithm has found parameter estimates that reasonably maximize the likelihood of observing the given data. In **Exhibit 1** we showcase how the convergence value throughout the different convergence criterion (Modified Relative Convergence MRC, Relative Convergence RC, and Absolute Convergence AC) decreases till epsilon ε. This indicates that the distance (absolute, relative or modified relative) between the old and new estimate decreases at each iteration, eventually satisfying the convergence criterion.

Exhibit 1: Convergence Values and Log-likelihood Values through iterations



Another diagnostic for convergence can be found by looking at the value of the log-likelihood of the estimates at each iteration. From iteration 1 to 7, as showcased by the graph on the right of **Exhibit 1,** the log-likelihood increases steadily. This indicates that the log likelihood of the vector of coefficient is ultimately maximised.

2. The Bayesian Metropolis Hastings Algorithm

The Bayesian Metropolis-Hastings Algorithm is a Markov Chain Monte Carlo method that iteratively generates and evaluates candidate samples to approximate posterior distributions in Bayesian inference, facilitating effective parameter estimation. It relies on a Metropolis-Hastings acceptance criterion for the acceptance or rejection of proposed samples.

Our implementation of this algorithm for a Probit regression takes 4 arguments:

- The dependent variable "Y", which is an array of binary responses
- The matrix of the explanatory variables "X", which is also an array
- The number of iterations n
- The proposal standard deviation of the normal distribution used to generate the proposal values for the parameter.

In implementing the Metropolis-Hastings algorithm for our Bayesian analysis, we have opted for a single-site update mechanism. Particularly, in each iteration, we update every coefficient, but just once at a time, corresponding to the coefficient to be updated, and we propose a new beta drawn from a normal distribution

centered on the previous coefficient. The reason we chose a single site update is to accommodate the different scaling of the intercept compared to the other covariates. This approach takes inspiration from the Gibbs Sampler, as in both cases we select just one parameter to update at a time. Another reason we did that was to allow for better exploration of the parameter space. Sometimes, a "good move" of one coefficient was discarded because of one coefficient being more impactful towards the dependent variable. The fact that it required more time for convergence was not really a problem, given the small dataset. Correlogram analysis of prior runs showed suboptimal convergence behavior of the intercept relative to other covariates. To solve this, we adjusted the standard deviation of the proposal distribution for the intercept to be an order of magnitude larger (specifically, 100 times greater) than that for other coefficients, set at 0.005.

We then compute the posterior probability and the log ratio, considering the following:

- We choose as a proposal distribution a normal distribution centered on the chosen coefficient;
- We chose as a prior distribution a standard normal distribution. The assumption seemed reasonable looking at the shape of the data, apart from some exceptions, like *age* or *tobacco*.

We then adopt the classic Metropolis Hastings framework: If the log-ratio is greater than or equal to a random number (sampled from a uniform of size 1), then the proposed parameter vector is accepted and the current beta is updated to the proposed beta.

2.1 Updating equations

To compute the likelihood ratio, we firstly compute $\mu_i = \Phi(x_i^T \beta)$. Given the Probit model we have at hand, we compute the likelihood as:

$$L(Y, X \mid \beta) = \prod_{i=1}^{n} [\mu^{y_i}(1-\mu)^{1-y_i}]$$

We take the logarithm to simplify computations:

$$\log L\,(Y, X \mid \beta) = \sum_{i=1}^{n} [y_i \log(\mu) + (1-y_i)\log(1-\mu)]$$

We compute the posterior probability(getting rid of constant values) for the current and the proposed vector of parameter as:

$$f(\beta|Y, X) = L(Y, X \mid \beta) \cdot \pi(\beta) = \sum_{i=1}^{n} [y_i \log(\mu) + (1-y_i)\log(1-\mu)] - \frac{\beta^2}{2}$$

Finally, we compute the likelihood log ratio;

$$\alpha(\beta, \beta') = f(\beta'|Y, X) - f(\beta|Y, X)$$

And we accept if: $U < \alpha(\beta, \beta')$, with U drawn from an Uniform(0,1).

2.2 Convergence diagnostic

To assess convergence, we use autocorrelation plots. We do not plot all of them here, although they can be found in the Notebook.

Exhibit 2:

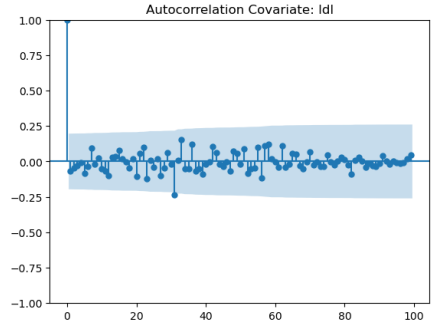


Exhibit 2: The autocorrelation coefficients fluctuate around zero with no clear pattern and remain within the confidence bands, suggesting that the *ldl* covariate does not exhibit significant autocorrelation. This indicates good mixing and suggests that the Markov chain for *ldl* may have converged to the true value.

Exhibit 3:

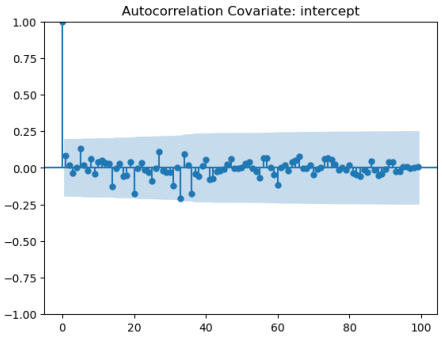


Exhibit 3: also the *intercept* chain exhibits low correlation, with similar reasoning regarding **Exhibit 2**. In previous versions, the autocorrelation was very persistent, and convergence was not attained.

3. Fitted models

We applied the two algorithms to the *SAheart* dataset to estimate the coefficients of a Probit regression with the dependent variable indicating the occurrence of a coronary heart disease(*chd*). Results are showcased in **Exhibit 4**.

Exhibit 4: Estimated coefficients

Variable	Fisher Scoring (FS) Alg.	Metropolis Hastings Alg.
intercept	-3.5700	-2.3825
sbp	0.0038	0.0011
tobacco	0.0482	0.0597
ldl	0.1030	0.0943
adiposity	0.0124	0.0466
typea	0.0235	0.0204
obesity	-0.0401	-0.0882
alcohol	1.955e-05	-0.0025
age	0.0263	0.0223
famhist	0.5390	0.3924

Exhibit 5: Significance of FS coefficients estimates

Variable	Standard error	z-score	p-value
intercept	0.751762	-4.749089	0.000002
sbp	0.003428	1.105448	0.268966
tobacco	0.015839	3.044441	0.002331
ldl	0.035289	2.913908	0.003569
adiposity	0.017382	0.713145	0.475756
typea	0.007188	3.277140	0.001049
obesity	0.026284	-1.527977	0.126518
alcohol	0.002686	0.007281	0.994191
age	0.007038	3.732720	0.000189
famhist	0.134819	3.997802	0.000064

The marginal effect of each covariate on the dependent variable for each statistical unit is obtained by computing the partial derivative:

$$\frac{\delta \Pr(Yi = 1|\, Xi)}{\delta\, Xi,j} = \frac{\delta \Phi(Xi^T \beta)}{\delta\, Xi,j} \beta_j = \beta_j \phi(Xi^T \beta)$$

where $\phi(Xi^T \beta)$ is the prime derivative of the cumulative distribution of a standard Gaussian distribution, which is its probability distribution.

Consequently, the marginal effect is not constant but depends both on the values of the other covariates and the starting value of the given covariate. This implies that each coefficient measures the direction of the effect (positive or negative) of each covariate on the probability that the individual suffered from a coronary heart disease, without providing information about the magnitude.

Based on the fitted models, it can be observed that systolic blood pressure, cumulative tobacco use, adiposity, Type-A behavior, age, and low-density lipoprotein cholesterol exhibit a positive impact on the dependent variable. This implies that a one-unit increase in each of these predictors results in an increase of the predicted probability of coronary heart disease. Moreover, the presence of a family history of heart disease increases the probability of coronary heart disease compared to those who do not have it. Contrary to expectations, obesity seems to have a negative effect on the dependent variable: if it increases by one unit, the predicted probability decreases. Finally, the current alcohol consumption appears to have no impact on the probability of suffering from coronary heart disease, as the estimated coefficients are close to zero in both algorithms.
However, these last two coefficients are not statistically significant at a 5% significance level, as the p-value is much higher. The same applies to adiposity, and systolic blood pressure. All the other coefficients are statistically

significant at a 1% significance level, and therefore, we reject the null hypothesis that they are equal to 0. The results are showcased in **Exhibit 5**, computed using the coefficient estimates of the Fisher Scoring algorithm.

Exhibit 6: Average Marginal Effects of Probit reg

Variable	Fisher Scoring Alg.	Metropolis Hastings Alg.
intercept	NA	NA
sbp	0.001093	0.000308
tobacco	0.013906	0.016955
ldl	0.029655	0.026789
adiposity	0.003575	0.013249
typea	0.006793	0.005801
obesity	-0.011582	-0.025086
alcohol	0.000006	-0.000706
age	0.007576	0.006349
famhist	0.155437	0.111493

The computation of Average Marginal Effects (AME) is useful to obtain insights into the magnitude of the effect of each variable. Using a probit regression, the average marginal effects represent the average of the marginal effects for each observation of the dataset. The marginal effect is the change in probability of the dependent variable taking value 1 (occurrence of a coronary heart disease) due to a one-unit change in one of the explanatory variables, assuming all others are kept constant. Based on the results showed in **Exhibit 6**, we identify four variables having a larger impact on the coronary heart disease:

- *famhist*, indicating if respondents have a family history of heart disease. The average marginal effects for this variable are slightly higher in the Fisher Scoring case but, in both algorithms, these are the largest. If the respondent has a family history of heart disease, assuming all other variables are kept constant, an occurrence of a coronary heart disease is 15.5% more likely using the Fisher Scoring algorithm or 11.1% more likely using the Metropolis Hastings algorithm.
- *ldl,* which is a measure of cholesterol. The average marginal effects for this variable are similar in both algorithms. Assuming all other variables are kept constant, a one-unit increase in low density lipoprotein cholesterol leads to a 3.0% increase in probability of coronary heart disease using the Fisher Scoring algorithm or a 2.7% increase using the Metropolis Hastings algorithm. This result matches the prior expectations about the high cholesterol – heart disease nexus. Scientific studies[1] in cardiovascular health literature confirm the connection between high LDL cholesterol levels and increased heart disease risk.
- *tobacco*, indicating the cumulative consumption of tobacco. Again, the average marginal effects for this variable are similar in both algorithms. Assuming all other variables are kept constant, a one-unit increase in the cumulative consumption of tobacco leads to a 1.4% increase in probability of coronary heart disease using the Fisher Scoring algorithm or a 1.7% increase using the Metropolis Hastings algorithm.
- *age*. Here the average marginal effects returned by the two algorithms differ just by 0.12%. A one unit increase in age leads to a 0.75% increase in the probability of coronary heart disease using the Fisher Scoring algorithm and or a 0.63% increase using the Metropolis Hastings algorithm.

1. Some studies were conducted by the Framingham Heart Study, an institute of research focused on cardiovascular disease. For more, https://www.framinghamheartstudy.org/fhs-about/