Probability and Statistics

Probit Regression

Gelsomino Ludovico[†]

[†]Bocconi University, Milan

January 30, 2024

I Introduction

There is a significant interest towards learning how the probability mass function of a binary response, denoted as $y \in \{0, 1\}$, changes with a set of observations $x = (x_1, ..., x_p)^T \in \mathbb{R}^p$. The nature of the problem precludes the use of common regression frameworks, such as linear regression, due to issues such as predicted probabilities outside $[0, 1]$, possible violations of constant variance and non-normality of the residuals. To overcome these challenges, a common approach is to consider the binary response y as a Bernoulli variable, where the probability parameter p is modelled as a linear combination of the predictors under a probit or logit mapping. Consequently, the problem is often reformulated as modeling the binary response through $\Pr(y = 1 \,|\, x, \beta) = \Phi(x^T\beta)$, with the objective of inferring the parameter vector $\beta = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ from the available data. In the subsequent sections, two distinct methods for estimating β will be presented. The first method follows a frequentist approach, utilizing Fisher Scoring for parameter estimation. The second method primarily employs Bayesian Statistics, leveraging data augmentation to obtain conjugate posteriors, which can be sampled using MCMC techniques such as Gibbs Sampler.

II Probit Regression using Fisher Scoring

Let $y = (y_1, \ldots, y_n)$ and let $X \in \mathbb{R}^{p\times n}$ be the corresponding design matrix whose generic row is $\mathbf{x}_i^T = (1, x_{i2}, \ldots, x_{ip})$, for $i = 1, \ldots, n$ [1]

The probit regression is usually formulated as the following Generalized Linear Model:

1. **Random Component:** $Y_i \overset{ind}{\sim} \text{Bern}(\mu_i)$
2. **Systematic Component:** $\eta_i = x_i^T\beta$
3. **Link Function:** $\eta_i = \Phi^{-1}(\mu_i)$

Therefore we have that, under probit regression $Y_i \overset{ind}{\sim} \text{Bern}(\Phi(x_i^T\beta))$. The corresponding pdf is

$$f(y_i\,; \Phi(x_i^T\beta)) = [\Phi(x_i^T\beta)]^{y_i}[1 - \Phi(x_i^T\beta)]^{(1-y_i)}$$

with

$$\mathbb{E}[Y_i] = \Phi(x_i^T\beta) \quad \text{and} \quad \mathbb{V}(Y_i) = \Phi(x_i^T\beta)[1 - \Phi(x_i^T\beta)]$$

In an ideal frequentist setting, we would infer the parameters β of the Regression analytically, by maximizing the Likelihood Function:

$$\hat{\beta}_{MLE} = \arg\max_{\beta} \mathcal{L}(\beta; Y, X)$$

where

$$\mathcal{L}(\beta; Y, X) = \prod_{i=1}^{n}[\Phi(x_i^T\beta)]^{y_i}[1 - \Phi(x_i^T\beta)]^{(1-y_i)}$$

Unfortunately, there is no readily available analytical solution for this specific likelihood function. Consequently, the application of an optimization technique becomes necessary.

[1] From now on I will denote the vector $\mathbf{x}_i$ with x_i

II.I The Fisher Scoring Algorithm for Probit Regression

In the case of a Generalized Linear Model (GLM) model, the updating equation of the Fisher scoring method is expressed as:

$$\beta_{t+1} = \beta_t + \mathbb{I}(\beta_t)^{-1} U(\beta_t)$$

where U denotes the vector of partial derivatives of the log-likelihood, and β_t is the vector of the current updates of the coefficients vector estimate. It is necessary to compute the elements of the vector $l'(\beta)$ and the Fisher information matrix.

For $Y = (Y_1, \ldots, Y_n)^T$ and $\mu = (\mu_1, \ldots, \mu_n)^T$, we can write the Fisher information matrix and the vector of partial-derivatives of the log-likelihood as :

$$\mathbb{I}(\beta) = X^T W X \quad , \quad U = X^T A (Y - \mu)$$

where W and A are diagonal matrix with elements

$$w_{ii} = \frac{1}{\mathbb{V}(Y_i)} \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^{-2}, \quad a_{ii} = \frac{1}{\mathbb{V}(Y_i)} \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^{-1}$$

Assuming to observe a vector $Y \overset{ind}{\sim} \text{Bern}(\Phi(X^T\beta))$ of n elements and a $p \times n$ covariate matrix, the update equation becomes

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

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

$$W_{ii} = \frac{1}{\mathbb{V}(Y_i)} \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^{-2} = \frac{1}{\mu_i(1-\mu_i)} \phi(\eta_i)^2 = \frac{\phi(x_i^T\beta)^2}{\Phi(x_i^T\beta)(1 - \Phi(x_i^T\beta))}$$

and Z is a $n \times 1$ vector with elements

$$Z_i = \eta_i + (Y_i - \mu_i) \left(\frac{\partial \eta_i}{\partial \mu_i} \right) = x_i^T \beta + \frac{Y_i - \Phi(x_i^T\beta)}{\phi(x_i^T\beta)}.$$

Algorithm 1 Probit Regression using Fisher Scoring

```
function Probit($X, \mathbf{y}, \epsilon$)
    Initialize $\beta^{(0)}$, $W^{(0)} \leftarrow I_{n\times n}$, $s \leftarrow 0$
    while $\frac{|\beta^{(s)} - \beta^{(s-1)}|}{|\beta^{(s)}| + \epsilon} \leq \epsilon$ do
        $\mu_i \leftarrow x_i^T \beta^{(s)}$                                    ▷ $i = 1, ..., n$
        $Z^{(s)} \leftarrow [z_1, ..., z_n]$ with $z_i = \mu_i + \frac{(Y_i - \Phi(\mu_i))}{\phi(\mu_i)}$
        $W_{j,j}^{(s)} \leftarrow \frac{\phi(\mu_j)^2}{\Phi(\mu_j)(1-\Phi(\mu_j))}$          ▷ $j = 1, ..., n$
        $\beta^{(s+1)} \leftarrow \left( X^T W_s X \right)^{-1} X^T W_s Z_s$
        $s \leftarrow s + 1$
    end while
    return $\beta^{(s)}$
end function
```

III Probit Regression: A Bayesian Approach

Let y be the data, following some distribution $\pi(y|\beta)$, i.e., the likelihood, with $\beta \in \Theta \subseteq \mathbb{R}^p$ being an unknown set of parameters. Let $\pi(\beta)$ be the prior distribution associated with β. In Bayesian analysis, inference is based on the posterior distribution for β, defined as

$$\pi(\beta|y) = \frac{\pi(\beta)\pi(y|\beta)}{\int_\Theta \pi(\beta)\pi(y|\beta)\, d\beta}.$$

Let $\mathbf{y} = (y_1, \ldots, y_n)^T$ be the vector of binary responses. Let X be the corresponding design matrix, with its generic row given by $x_i^T = (1, x_{i2}, \ldots, x_{ip})$, for $i = 1, \ldots, n$. We assume independent Gaussian priors $N(0, \sigma_0^2)$ for each $\beta, \ldots, \beta$. Hence,

$$\beta_j \sim N(0, \sigma_0^2) \quad j = 1, \ldots, p \quad \Rightarrow \pi(\beta) = \prod_{j=i}^{p} \phi(\beta_j; 0, \sigma_0^2)$$

The likelihood function of a probit regression model is as follows:

$$\begin{aligned}\pi(y|\beta) &= \prod_{i=1}^{n} [\Phi(x_i^T\beta))_i]^{y_i} \cdot [1 - \Phi(x_i^T\beta)]^{1-y_i} \\ &= \prod_{i=1}^{n} [\mathbb{1}(y_i = 1)\Phi(x_i^T\beta) + \mathbb{1}(y_i = 0)\{1 - \Phi(x_i^T\beta)\}]\end{aligned}$$

The posterior can be obtained via Bayes' rule. Under Gaussian priors, this yields:

$$\pi(\beta|y) = \frac{\prod_{j=1}^{p} \phi(\beta_j; 0, \sigma_0^2) \prod_{i=1}^{n} [\Phi(x_i^T\beta)]^{y_i} [1 - \Phi(x_i^T\beta)]^{(1-y_i)}}{\int_{\mathbb{R}^p} \prod_{j=1}^{p} \phi(\beta_j; 0, \sigma_0^2) \prod_{i=1}^{n} [\Phi(x_i^T\beta)]^{y_i} [1 - \Phi(x_i^T\beta)]^{(1-y_i)} d\beta}$$

whose normalizing constant is often intractable and difficult to compute.

Latent Variable Representation

To ease the computation, we introduce a vector of latent variables $z \in Z \subseteq \mathbb{R}^q$ and we treat them as if they were an additional set of unknown parameters. We let $\pi(y, z|\beta)$ be the augmented likelihood and we define the augmented posterior

$$\pi(\beta, z|y) \propto \pi(y, z|\beta)\pi(\beta)$$

In this framework, we focus on the augmented posterior $\pi(\beta, z|y)$, in the hope that it is more tractable than $\pi(\beta|y)$ under suitable data augmentation strategies. Usually, these strategies lead to conditional conjugacy results for $\pi(\beta|y, z)$ and $\pi(z|y, \beta)$. The conditional conjugacy allows, for example, easier sampling from $\pi(\beta, z|y)$ relative to directly targeting $\pi(\beta|y)$ because the augmented likelihood is typically more tractable than the original one. If one is interested only in the original parameters β or in the latent dimensions z, then it suffices to ignore the other set of sampled parameters.

Latent Variable Gibbs Sampler

If the full conditionals are available in closed form, we can apply a Gibbs Sampling strategy to target the augmented posterior $\pi(\beta, z|y)$.

Step 1: Sample from the 'posterior' of β based on the complete likelihood $\pi(\beta|z, y, X)$

Step 2: Sample z from the full conditional $\pi(z|\beta, y, X)$

Sampling iteratively from the distribution in step 1 and step 2 yields a Markov chain that is guaranteed to have as stationary distribution the actual posterior $\pi(z, \beta|y, X)$. Hence, samples from β correspond to those from $\pi(\beta|y, X)$.

Gibbs Sampler with Uninformative Prior

Let $z = (z_i, ..., z_n)$ with $z_i \sim \mathcal{N}(x_i^T\beta, 1)$ be a vector of latent variables and set our binary variable $y_i = \mathbb{1}(z_i > 0)$ for $i = 1, ..., n$. Assuming constant priors, we get $\pi(\beta) \propto 1$.

In order to apply a Gibbs Sampling strategy, we need to retrieve the full conditionals

$$\beta|z, y, X \text{ and } z|\beta, y, X$$

Deriving the full conditional of β

Given that β is conditionally independent of $y|z$, we get $\pi(\beta|z, y, X) = \pi(\beta|z, X)$ and the augmented posterior can be reduced to:

$$\pi(\beta, z|y) \propto \pi(y, z|\beta)\pi(\beta)$$

Thus

$$\begin{aligned}\pi(\beta|z, X) &\propto \pi(\beta)\pi(z|\beta, X)\\ &\propto \exp\left\{-0.5(z - X\beta)^T(z - X\beta)\right\}\\ &\propto \exp\left\{-0.5(\beta - (X^TX)^{-1}X^Tz)^T(X^TX)(\beta - (X^TX)^{-1}X^Tz)\right\}\end{aligned}$$

Thus, under the constant prior for the coefficients, the full conditional β is known:

$$\beta|z, X \sim \mathcal{N}_p(\mu, \Sigma), \quad \mu = \Sigma X^Tz, \quad \Sigma = (X^TX)^{-1}.$$

Deriving the full conditional of z

To "impute" data z, recall that each $\pi(z_i|\beta, X) = \phi(z_i; x_i^T\beta, 1)$ independently for each $i = 1, \ldots, n$. Hence, $\pi(z|y, \beta) = \prod_{i=1}^n \pi(z_i|y_i, \beta)$, where each term is proportional to $\phi(z_i; x_i^T\beta, 1) \cdot \mathbb{1}(z_i > 0)$ if $y_i = 1$ or to $\phi(z_i; x_i^T\beta, 1) \cdot \mathbb{1}(z_i \leq 0)$ if $y_i = 0$. This means that each $\pi(z_i|y_i, \beta)$ is a truncated normal distribution

$$z_i|\beta, y, x_i \sim \begin{cases}\mathcal{TN}(x_i^T\beta, 1, 0, \infty) & \text{if } y_i = 1\\ \mathcal{TN}(x_i^T\beta, 1, -\infty, 0) & \text{if } y_i = 0\end{cases}$$

Gibbs Sampling strategy based on full conditionals with constant priors

Step 1: Sample $z_{1,\ldots,n}^{(s)} \sim \begin{cases}\mathcal{TN}(x_i^T\beta^{(s-1)}, 1, 0, \infty) & \text{if } y_i = 1\\ \mathcal{TN}(x_i^T\beta^{(s-1)}, 1, -\infty, 0) & \text{if } y_i = 0\end{cases}$

Step 2: Sample $\beta^{(s)} \sim \mathcal{N}_p(\mu, \Sigma)$ with $\mu = \Sigma X^Tz^{(s)},$ and $\Sigma = (X^TX)^{-1}$

Gibbs Sampler with Independent Gaussian Priors

Let $z = (z_i, ..., z_n)$ with $z_i \sim \mathcal{N}(x_i^T\beta, 1)$ be a vector of latent variables and set our binary variable $y_i = \mathbb{1}(z_i > 0)$ for $i = 1, ..., n$. Assuming independent Gaussian priors $\beta_j \sim N(0, \sigma_0^2) \quad j = 1, \ldots, p$ we get

$$\pi(\beta) = \prod_{j=i}^{p} \phi(\beta_j; 0, \sigma_0^2)$$

In order to apply a Gibbs Sampling strategy, we need to retrieve the full conditionals

$$\beta|z, y, X \text{ and } z|\beta, y, X$$

Deriving the full conditional of β

Given that β is conditionally independent of $y|z$, we get $\pi(\beta|z, y, X) = \pi(\beta|z, X)$ and the augmented posterior can be reduced to:

$$\pi(\beta, z|y) \propto \pi(y, z|\beta)\pi(\beta)$$

Thus

$$\begin{aligned}\pi(\beta|z, X) &\propto \pi(\beta)\pi(z|\beta, X)\\ &\propto \exp\left\{-0.5\beta^T(\sigma^{-2}I_p)\beta\right\}\exp\left\{-0.5(z - X\beta)^T(z - X\beta)\right\}\\ &\propto \exp\left\{-0.5\beta^T(X^TX + \sigma^{-2}I_p)\beta + \beta^TX^Tz\right\}\end{aligned}$$

Calling $\Sigma = (X^TX + \sigma^{-2}I_p)^{-1}$

$$\begin{aligned}\pi(\beta|z, X) &\propto \exp\left\{-0.5\beta^T\Sigma^{-1}\beta + \beta^T\Sigma^{-1}\Sigma X^Tz - 0.5z^TX\Sigma\Sigma^{-1}\Sigma X^Tz\right\}\\ &\propto \exp\left\{-0.5(\beta - \Sigma X^Tz)^T\Sigma^{-1}(\beta - \Sigma X^Tz)\right\}\end{aligned}$$

Thus, under Gaussian priors for the coefficients, the full conditional β is known:

$$\beta|z, X \sim \mathcal{N}_p(\mu, \Sigma), \quad \mu = \Sigma X^Tz, \quad \Sigma = (X^TX + \sigma_0^{-2}I_p)^{-1}.$$

Deriving the full conditional of z

To "impute" data z, recall that each $\pi(z_i|\beta, X) = \phi(z_i; x_i^T\beta, 1)$ independently for each $i = 1, \ldots, n$. Hence, $\pi(z|y, \beta) = \prod_{i=1}^n \pi(z_i|y_i, \beta)$, where each term is proportional to $\phi(z_i; x_i^T\beta, 1) \cdot \mathbb{1}(z_i > 0)$ if $y_i = 1$ or to $\phi(z_i; x_i^T\beta, 1) \cdot \mathbb{1}(z_i \leq 0)$ if $y_i = 0$. This means that each $\pi(z_i|y_i, \beta)$ is a truncated normal distribution

$$z_i|\beta, y, x_i \sim \begin{cases}\mathcal{TN}(x_i^T\beta, 1, 0, \infty) & \text{if } y_i = 1\\ \mathcal{TN}(x_i^T\beta, 1, -\infty, 0) & \text{if } y_i = 0\end{cases}$$

Gibbs Sampling strategy based on full conditionals with Independent Gaussian priors

Step 1: Sample $z_{1,\ldots,n}^{(s)} \sim \begin{cases}\mathcal{TN}(x_i^T\beta^{(s-1)}, 1, 0, \infty) & \text{if } y_i = 1\\ \mathcal{TN}(x_i^T\beta^{(s-1)}, 1, -\infty, 0) & \text{if } y_i = 0\end{cases}$

Step 2: Sample $\beta^{(s)} \sim \mathcal{N}_p(\mu, \Sigma)$ with $\mu = \Sigma X^Tz^{(s)}$, and $\Sigma = (X^TX + \sigma_0^{-2}I_p)^{-1}$

Algorithm 2 Bayes Probit MCMC (Uninformative Prior)

```
function BayesProbitUP($X, \mathbf{y}, S$)
    Initialize $\beta^{(0)}, z^{(0)}$
    Beta Chain ← $S \times p$ Matrix                    ▷ Matrix containing the $\beta$ MCMC
    $\Sigma \leftarrow (X^T X)^{-1}$
    for $s \leftarrow 1$ to $S$ do
        Sample $z^{(s)}_{1,\ldots,n}$ from $\sim \begin{cases} \mathcal{TN}(x_i^T \beta^{(s-1)}, 1, 0, \infty) & \text{if } y_i = 1 \\ \mathcal{TN}(x_i^T \beta^{(s-1)}, 1, -\infty, 0) & \text{if } y_i = 0 \end{cases}$
        $\mu \leftarrow \Sigma X^T z$
        Sample $\beta^{(s)}$ from $\mathcal{N}_p(\mu, \Sigma)$
        Beta Chain$[s, :] \leftarrow \beta^{(s)}$           ▷ Insert $\beta^{(s)}$ in row s of Beta Chain
    end for
    return Beta Chain
end function
```

Algorithm 3 Bayes Probit MCMC (Independent Gaussian Priors)

```
function BayesProbit($X, \mathbf{y}, S$)
    Initialize $\beta^{(0)}, z^{(0)}, \sigma_0$
    Beta Chain ← $S \times p$ Matrix                    ▷ Matrix containing the $\beta$ MCMC
    $\Sigma \leftarrow ({\sigma_0}^{-2} + X^T X)^{-1}$
    for $s \leftarrow 1$ to $S$ do
        Sample $z^{(s)}_{1,\ldots,n}$ from $\sim \begin{cases} \mathcal{TN}(x_i^T \beta^{(s-1)}, 1, 0, \infty) & \text{if } y_i = 1 \\ \mathcal{TN}(x_i^T \beta^{(s-1)}, 1, -\infty, 0) & \text{if } y_i = 0 \end{cases}$
        $\mu \leftarrow \Sigma X^T z$
        Sample $\beta^{(s)}$ from $\mathcal{N}_p(\mu, \Sigma)$
        Beta Chain$[s, :] \leftarrow \beta^{(s)}$           ▷ Insert $\beta^{(s)}$ in row s of Beta Chain
    end for
    return Beta Chain
end function
```

IV Results & Conclusions

Table 1: Coefficients with Different Priors

Variable	Fisher Scoring	Gibbs Sampl. Unif. Prior	Gibbs Sampl. Gauss. Prior
const	(3.5702)	(3.3349)	(3.319)
sbp	0.0038	0.0034	0.0034
tobacco	0.0482	0.0444	0.0444
ldl	0.1028	0.0934	0.0933
adiposity	0.0124	0.0112	0.0113
famhist	0.539	0.4895	0.4893
typea	0.0236	0.0219	0.0218
obesity	(0.0402)	(0.0359)	(0.0362)
alcohol	0.0	0.0002	0.0002
age	0.0263	0.0249	0.0248

As shown from table 1, the results obtained from the Fisher scoring, Bayesian Probit with Uniform Prior and Bayesian Probit with Independent Gaussian Priors are quite similar [2]. Unfortunately, the interpretation of the estimated coefficients of a probit regression is not as straightforward as for classical linear regression.

To ease the interpretation, we will calculate, starting from the coefficients, the average marginal effects using:

$$\text{AME}_{\beta_j} = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i^T\beta)\beta_j$$

for continuous variables and

$$\begin{aligned}\text{AME}_{\beta_{famhist}} = \frac{1}{n}\sum_{i=1}^{n}[\Phi(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 + ... + \beta_j x_j) \\ - \Phi(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + ... + \beta_j x_j)] \quad j \in \{1, ..., n\}\end{aligned}$$

for the binary variable famhist.

Table 2: Average Marginal Effects

const	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age
(0.9926)	0.001	0.0133	0.0279	0.0034	0.138	0.0065	(0.0108)	0.0000	0.0074

Based on the provided marginal effects, assuming the statistical significance of the variables, we observe predominantly positive impacts on the disease probability for most coefficients, except for obesity and alcohol (which has no effects). Notably, family history and LDL (cholesterol) exhibit the most substantial effects on the probability of disease. Specifically, a one-unit rise in cholesterol corresponds to a 3% increase (on average) in the disease probability and the presence of a family history elevates the disease probability by ≈ 15% (on average).

[2] See [Appendix] for algorithmic convergence

Appendix

Diagnostic on β

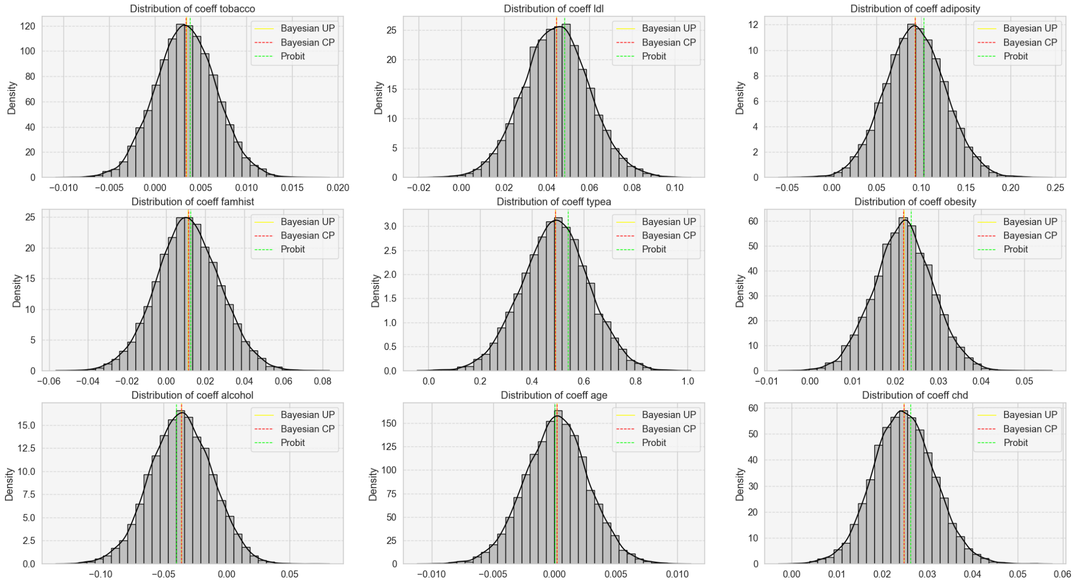

Figure 1: Distribution of Coefficients

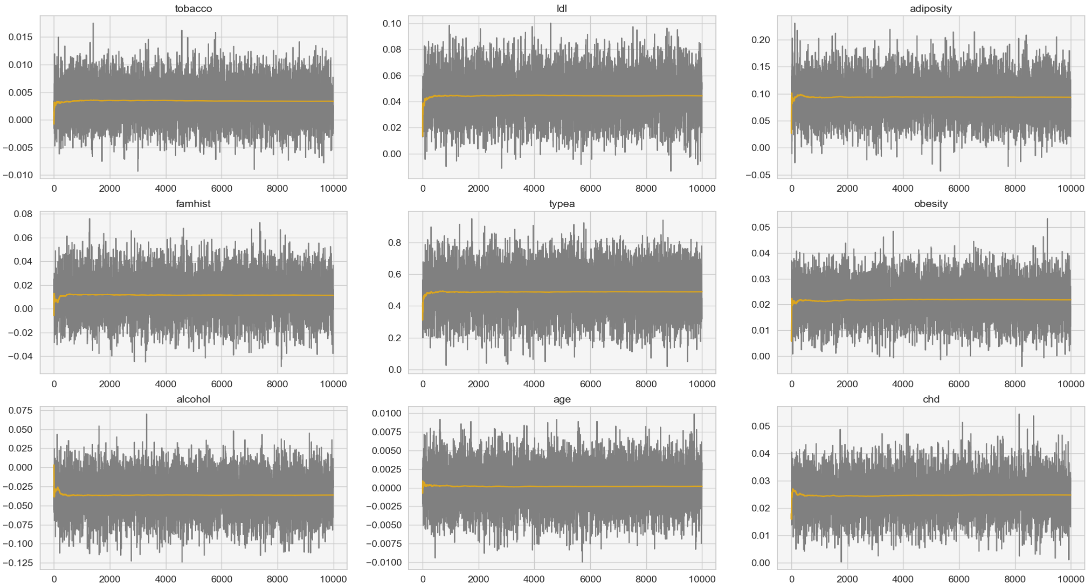

Figure 2: Trace Plots

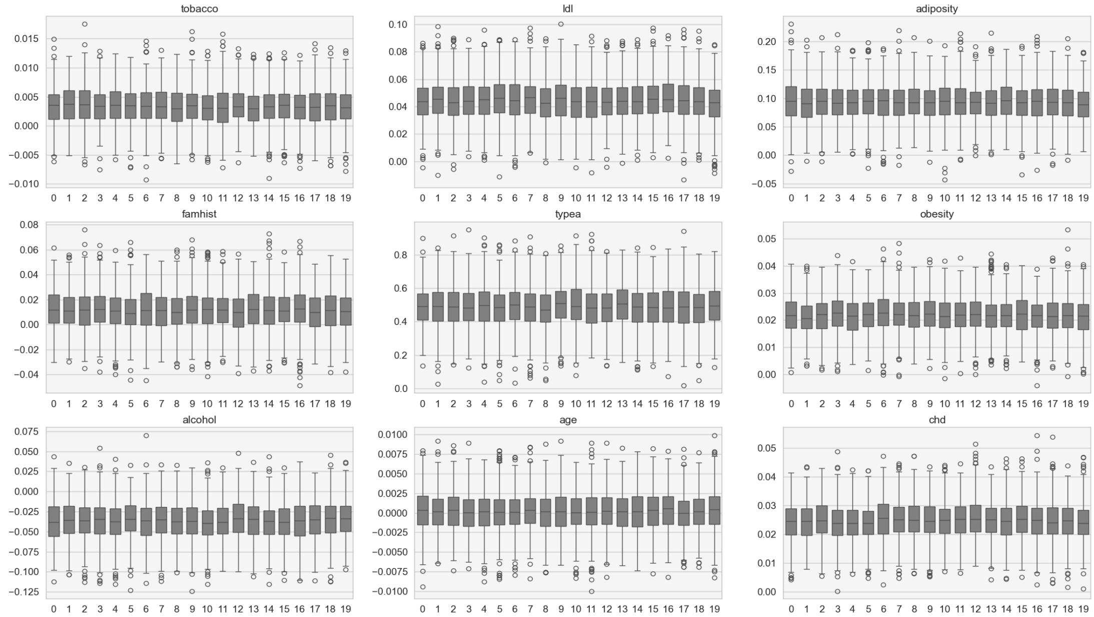

Figure 3: Trace Plots

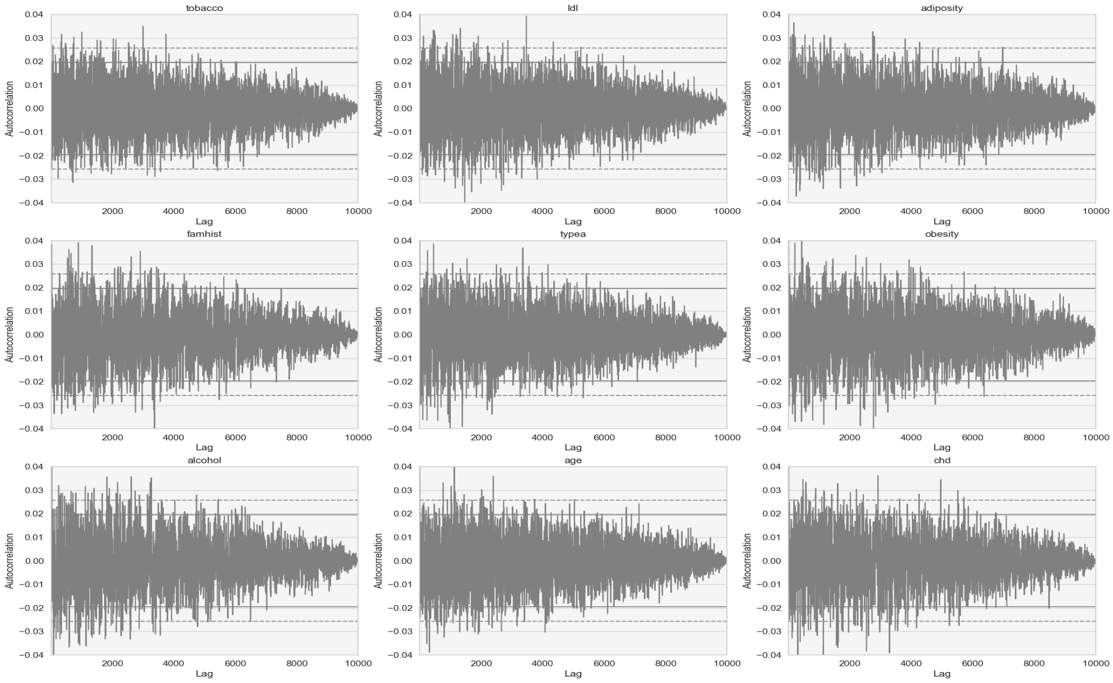

Figure 4: Autocorrelation Plots

Diagnostic on z

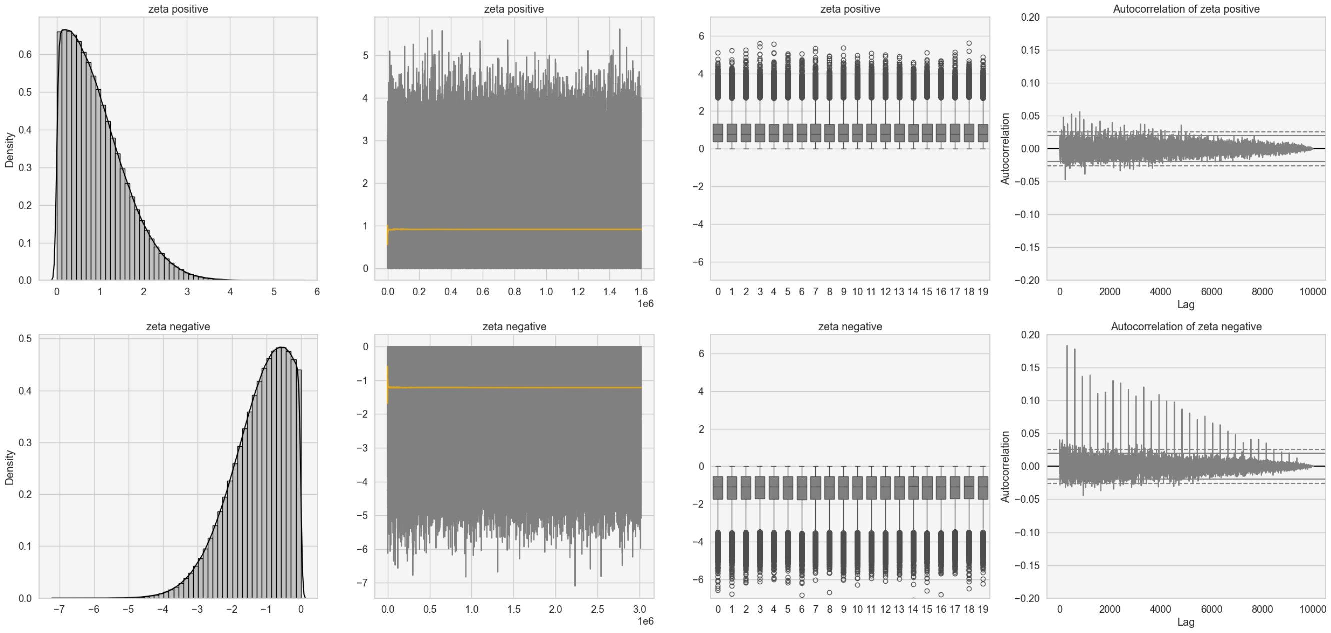

Figure 5: Diagnostic on z