

Bayesian analysis for male fertility

Enrico Grimaldi
1884443

Statistical Methods for Data Science 2

La Sapienza University of Rome
a.y. 2022/2023

Final project for the course:

Contents

1. Introduction	3
2. Exploratory data analysis	4
2.1 Overview of the data set	4
2.2 Plot and visualize	7
2.3 Study of the correlations	9
3. Proposed statistical models	9
3.1 Bayesian logistic regression model	9
3.2 Bayesian probit regression model	10
4. First model	11
4.1 MCMC convergence analysis	12
4.2 Prediction	12
5. Second model	13

1. Introduction

In Italy, the issue of fertility is quite central in recent years, especially in the post-covid period, there has even been talk of a recessionary phase in terms of birth rates. The underlying causes are many and particularly diverse. For example, among the causes of the decline in first children is the prolonged stay of young people in the family of origin, which in turn is due to multiple factors: the protracted length of time in education, the difficulties young people face in entering the world of work and the widespread instability of work itself, difficulties in accessing the housing market, a long-term trend of low economic growth, as well as other possible factors of a cultural nature.

Moreover, Italy is a country with a high level of emigration and a tendency on the part of the younger age group to seek more prosperous opportunities abroad, leading to a lower average age.

In addition to this purely socio-economic reasons and the crisis the country is experiencing in this regard, another rather worrying trend was found to be strongly affecting the birth rate in most Western countries: an analysis of 56 countries from 6 different continents recorded a halving in sperm count from 1973 to 2018.

The fertility status of men affects the issue introduced above in a worrisome way and needs more attention and awareness.

The aim of our analysis is to study the main sources of infertility for a man. Given a data set with a set of attributes deemed more or less informative, we choose to use a simple Bayesian logistic regression for the prediction of subject with normal or altered sperm (analyzed according to the WHO (2010) criteria).

What we seek to highlight through this study is the relationship between bad habits, health status, context (environment, social, and time of year), and fertility level.

Keep in mind, however, that the real goal of the project is not to achieve high prediction performance but to analyze the model parameters and infer how individual covariates affect the (binary) label value.

Let's give a look to the first rows of the used data set from the UCI repository:

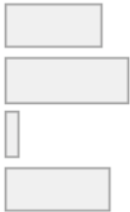
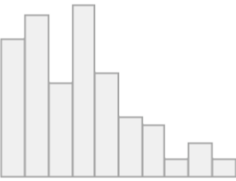
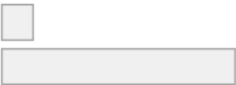
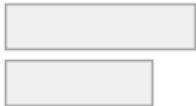

Season	Age	Disease	Accident	Surgery	Fever	Alcohol	Smoking	Sedentarity	Out
-0.33	0.94	1	0	1	0	0.8	1	0.31	O
-0.33	0.50	1	0	0	0	1.0	-1	0.50	N
-0.33	0.75	0	1	1	0	1.0	-1	0.38	N
-0.33	0.67	1	1	0	0	0.8	-1	0.50	O
-0.33	0.67	1	0	1	0	0.8	0	0.50	N
-0.33	0.67	0	0	0	-1	0.8	-1	0.44	N

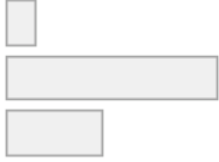
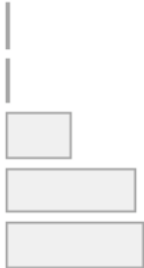

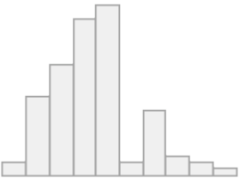
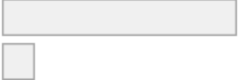
2. Exploratory data analysis

2.1 Overview of the data set

Dimensions: 99 x 10

Duplicates: 0

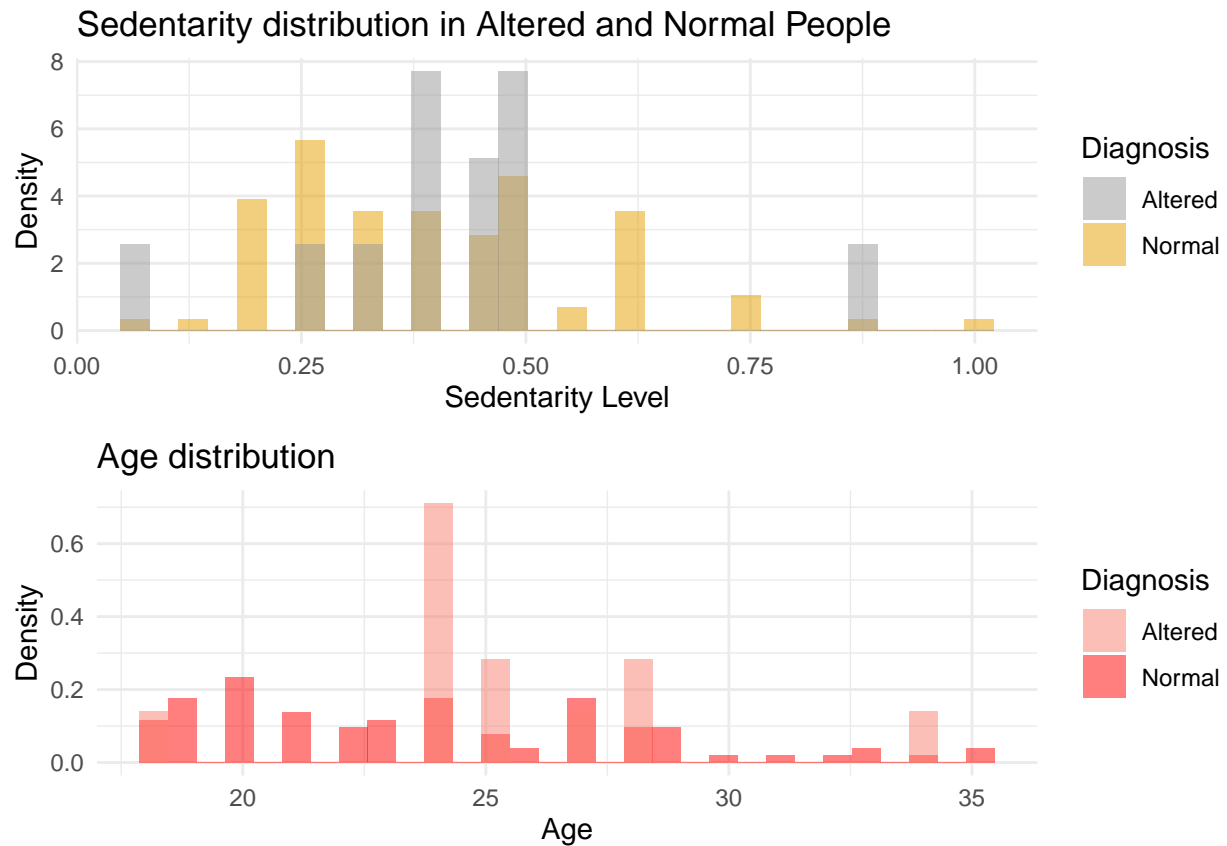
Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
Season [numeric]	Mean (sd) : -0.1 (0.8) min < med < max: -1 < -0.3 < 1 IQR (CV) : 2 (-10.5)	-1.00 : 28 (28.3%) -0.33 : 36 (36.4%) 0.33 : 4 (4.0%) 1.00 : 31 (31.3%)		0 (0.0%)
Age [numeric]	Mean (sd) : 0.7 (0.1) min < med < max: 0.5 < 0.7 < 1 IQR (CV) : 0.2 (0.2)	18 distinct values		0 (0.0%)
Disease [integer]	Min : 0 Mean : 0.9 Max : 1	0 : 12 (12.1%) 1 : 87 (87.9%)		0 (0.0%)
Accident [integer]	Min : 0 Mean : 0.4 Max : 1	0 : 56 (56.6%) 1 : 43 (43.4%)		0 (0.0%)
Surgery [integer]	Min : 0 Mean : 0.5 Max : 1	0 : 49 (49.5%) 1 : 50 (50.5%)		0 (0.0%)

Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
Fever [integer]	Mean (sd) : 0.2 (0.6) min < med < max: -1 < 0 < 1 IQR (CV) : 1 (3)	-1 : 9 (9.1%) 0 : 62 (62.6%) 1 : 28 (28.3%)		0 (0.0%)
Alcohol [numeric]	Mean (sd) : 0.8 (0.2) min < med < max: 0.2 < 0.8 < 1 IQR (CV) : 0.2 (0.2)	0.20 : 1 (1.0%) 0.40 : 1 (1.0%) 0.60 : 19 (19.2%) 0.80 : 38 (38.4%) 1.00 : 40 (40.4%)		0 (0.0%)
Smoking [integer]	Mean (sd) : -0.4 (0.8) min < med < max: -1 < -1 < 1 IQR (CV) : 1 (-2.3)	-1 : 56 (56.6%) 0 : 22 (22.2%) 1 : 21 (21.2%)		0 (0.0%)
Sedentarity [numeric]	Mean (sd) : 0.4 (0.2) min < med < max: 0.1 < 0.4 < 1 IQR (CV) : 0.2 (0.5)	14 distinct values		0 (0.0%)
Out [character]	1. N 2. O	87 (87.9%) 12 (12.1%)		0 (0.0%)

As can be noted from the table above (100×10) the data set is mostly characterized by categorical variables hashed into numeric (discrete) variables and to a lesser extent by discrete range-limited variables. In each case there are no missing values for any attribute and the values that can be assumed have a very specific meaning as below:

- **Season** \rightarrow season in which the analysis was performed:
 - winter = -1;
 - spring = -0.33;
 - summer = 0.33;
 - fall = 1.
- **Age** \rightarrow age at the time of analysis normalized from (18,36) to (0, 1)
- **Disease** \rightarrow childish diseases (i.e., chicken pox, measles, mumps, polio) yes/no (0, 1)
- **Accident** \rightarrow accident or serious trauma, yes/no \rightarrow (0, 1)
- **Surgery** \rightarrow surgical intervention, yes/no \rightarrow (0, 1)
- **Fever** \rightarrow high fevers in the last year:
 - less than three months ago = -1;
 - more than three months ago = 0;
 - no = 1.
- **Alcohol** \rightarrow frequency of alcohol consumption (quantized in 5 numbers $\in (0, 1)$):
 - several times a day;
 - every day;
 - several times a week;
 - once a week;
 - hardly ever or never.
- **Smoking** \rightarrow smoking habit:
 - never = 1;
 - occasional = 0;
 - daily = 1.
- **Sedentarity** \rightarrow number of hours spent sitting per day (normalized to 16 hours)
- **Out** \rightarrow diagnosis:
 - normal (N);
 - altered (O).

2.2 Plot and visualize



First, let's try to analyze two distributions by distinguishing them for subjects we know have "Altered or"Normal" sperm counts:

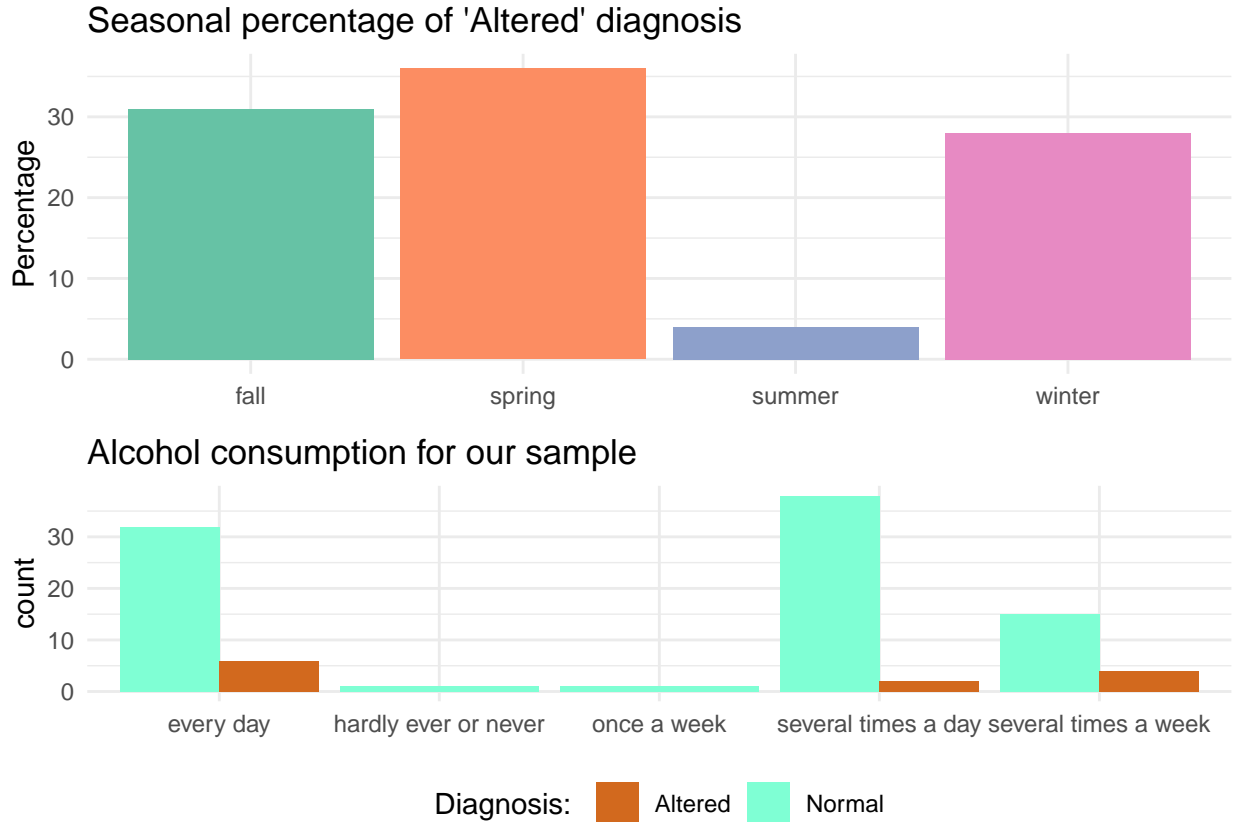
1. *Sedentarity distribution*: in both cases it seems to behave as a normal (or similar) distribution concentrated around a mean value;
2. *Age distribution*: unlike the previous case it seems that the two populations show different peaks and in particular the "Normal" population shows a rather flat pattern.

We chose to focus on these distributions first because they are characterized by a wide range of values relative to the rest of the other variables, and to provide an initial perspective on how aspects of the individual, both in terms of habits (sedentarity) and in terms of intrinsic characteristics (age), may exhibit similar patterns.

The main inferences we can draw from these plots are that:

- The "Normal" population has smaller peaks in its concentration (but this difference could also be due to the scarcity of samples for "Altered" diagnoses).

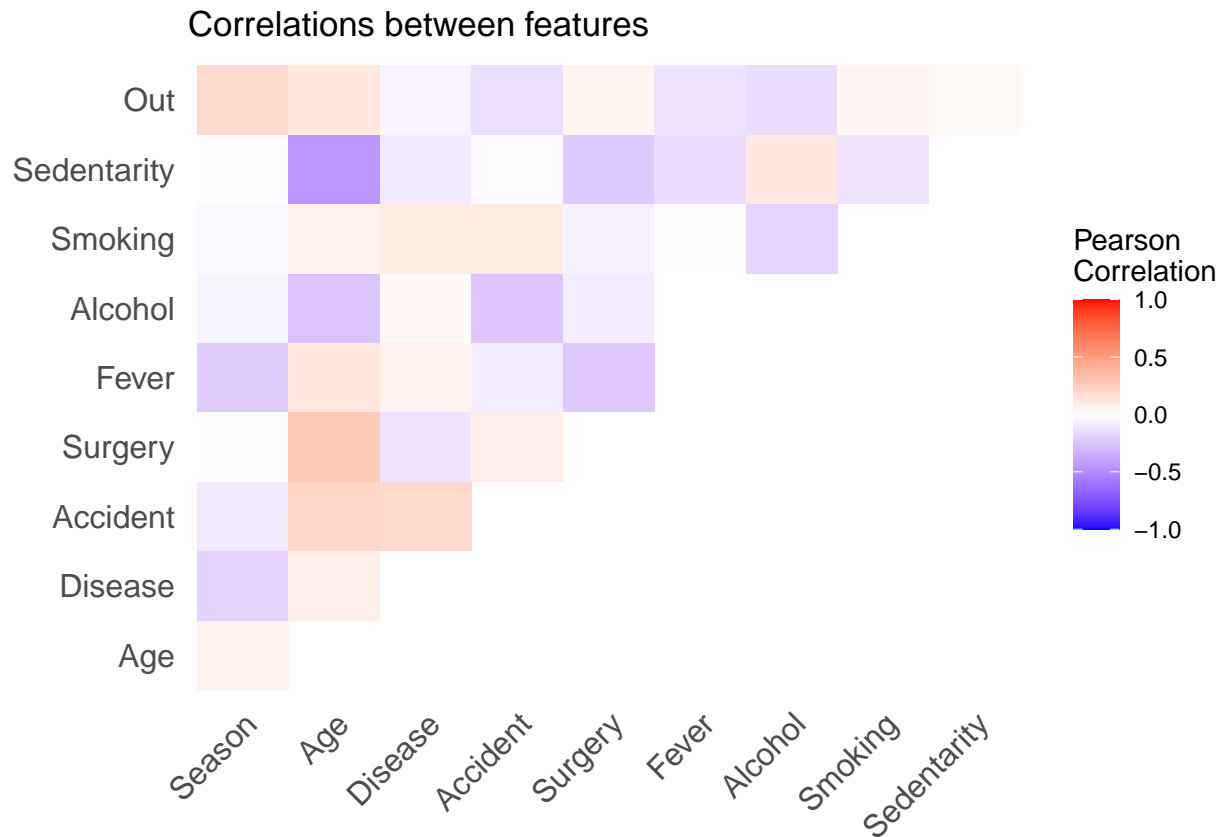
- Very high levels of sedentariness and older age (than the considered average) are not necessarily a cause of infertility but may influence it. It should be remembered in particular that the age range was limited and one possible explanation for the peak around age 25 is that those years generally see a certain group of subjects facing important periods of stress as they transition to the world of work.



As a second interesting analysis we report how two very different variables influence the number (or percentage) of “Altered” subjects:

- the variable “*Season*” directly related to the external environment and decisive (probably also due to other hidden variables) in defining the class of our sample;
- the variable “*Alcohol*” linked again to the habits of the subject and considered by the scientific literature to be of relevant importance in our classification problem.

2.3 Study of the correlations



3. Proposed statistical models

Given the binary nature of the output (variable “Diagnosis”/“Out”) it is considered valid to consider a model that is as simple as it is effective: a *Bayesian logistic regression*. This model will be used to classify subjects into “Altered” or “Normal” but the real goal is to study the estimation of the mean value of our output variable as a composition of the parameters of interest.

3.1 Bayesian logistic regression model

The Bayesian logistic regression model defines the distribution of the output random variable as a Bernoulli of parameter p (mean of the distribution), in particular implies the following parameterization and prior distributions of the parameters:

$$\begin{aligned}
Y|p &\sim \text{Ber}(p) \\
E[Y|p] &= p \\
\text{logit}(p) &= \beta_{Seas} \cdot X_1 + \beta_{Age} \cdot X_2 + \beta_{Dis} \cdot X_3 + \beta_{Acc} \cdot X_4 + \beta_{Surg} \cdot X_5 + \beta_{Al} \cdot X_6 + \\
&\quad + \beta_{Fev} \cdot X_7 + \beta_{Smok} \cdot X_8 + \beta_{Sed} \cdot X_9
\end{aligned}$$

The logistic regression coefficients β associated with a predictor X represents the relationship between the output variable and a specific feature: it is the expected change in log odds of having the outcome per unit change in X . So increasing the predictor by 1 unit (or going from 1 level to the next) multiplies the odds of having the outcome by e^β .

$$\begin{aligned}
\beta_{Seas} &\sim N(0, \sigma_{Seas}) \\
\beta_{Age} &\sim N(0, \sigma_{Age}) \\
\beta_{Dis} &\sim N(0, \sigma_{Dis}) \\
\beta_{Acc} &\sim N(0, \sigma_{Acc}) \\
\beta_{Surg} &\sim N(0, \sigma_{Surg}) \\
\beta_{Al} &\sim N(0, \sigma_{Al}) \\
\beta_{Fev} &\sim N(0, \sigma_{Fev}) \\
\beta_{Smok} &\sim N(0, \sigma_{Smok}) \\
\beta_{Sed} &\sim N(0, \sigma_{Sed})
\end{aligned}$$

Summarizing then, a prior was defined for each parameter (each related to a specific feature of the data set), and the mean p of our Bernoulli variable in output Y is defined as the following transformation of the above parameters:

$$P(Y = 1) = p = \frac{1}{1 + e^{\sum_i \beta_i X_i}}$$

To predict the fertility status (binary) of a man we will then use the logistic function (inverse of logit), obtaining a likelihood in view of our prior beliefs about the parameters.

3.2 Bayesian probit regression model

The Probit model is a statistical method used for binary classification and estimation of probabilities. It's particularly useful when dealing with categorical outcomes where the response variable can take one of two possible values, often denoted as 0 or 1. The model assumes that there is an underlying continuous latent variable that determines the outcome, and this latent variable follows a standard normal (Gaussian) distribution.

In the Probit model, the relationship between the predictor variables (also called independent variables or features) and the binary outcome is modeled through the cumulative distribution function of the standard normal distribution. This function transforms the linear combination of predictor variables into a probability that the outcome will be 1.

Mathematically, the Probit model can be represented as follows:

$$P(Y = 1|X) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)$$

Where:

$P(Y = 1|X)$ is the probability of the binary outcome being 1 given the predictor variables X . Φ represents the cumulative distribution function of the standard normal distribution. $\beta_0, \beta_1, \dots, \beta_k$ are the coefficients associated with the predictor variables X_1, X_2, \dots, X_k . The Probit model estimates these coefficients based on the observed data, aiming to find the best-fitting model that explains the relationship between the predictors and the binary outcome. The coefficients indicate how the predictor variables influence the probability of the binary outcome occurring.

In summary, the Probit model is a statistical approach used to analyze and predict binary outcomes by modeling the relationship between predictor variables and probabilities through the cumulative distribution function of the standard normal distribution.

4. First model

The first model then consists of a simple logistic regression in which the parameters used are relative to each of the variables in the original data set.

The model is written with JAGS (just another gibbs sampler), which allows us to easily pursue sampling methods based on Monte Carlo Markov Chains.

```
cat("model {
  for (i in 1:N) {
    Y[i] ~ dbern(p[i])
    logit(p[i]) <- beta0 + beta1*Season[i] + beta2*Age[i]
    + beta3*Disease[i] + beta4*Accident[i] + beta5*Surgery[i]
    + beta6*Fever[i] + beta7*Alcohol[i] + beta8*Smoking[i]
    + beta9*Sedentarity[i]
  }

  # Priors
  beta0 ~ dnorm(mu_0, sig_0)
  beta1 ~ dnorm(mu_0, sig_0)
  beta2 ~ dnorm(mu_0, sig_0)
  beta3 ~ dnorm(mu_0, sig_0)
```

```

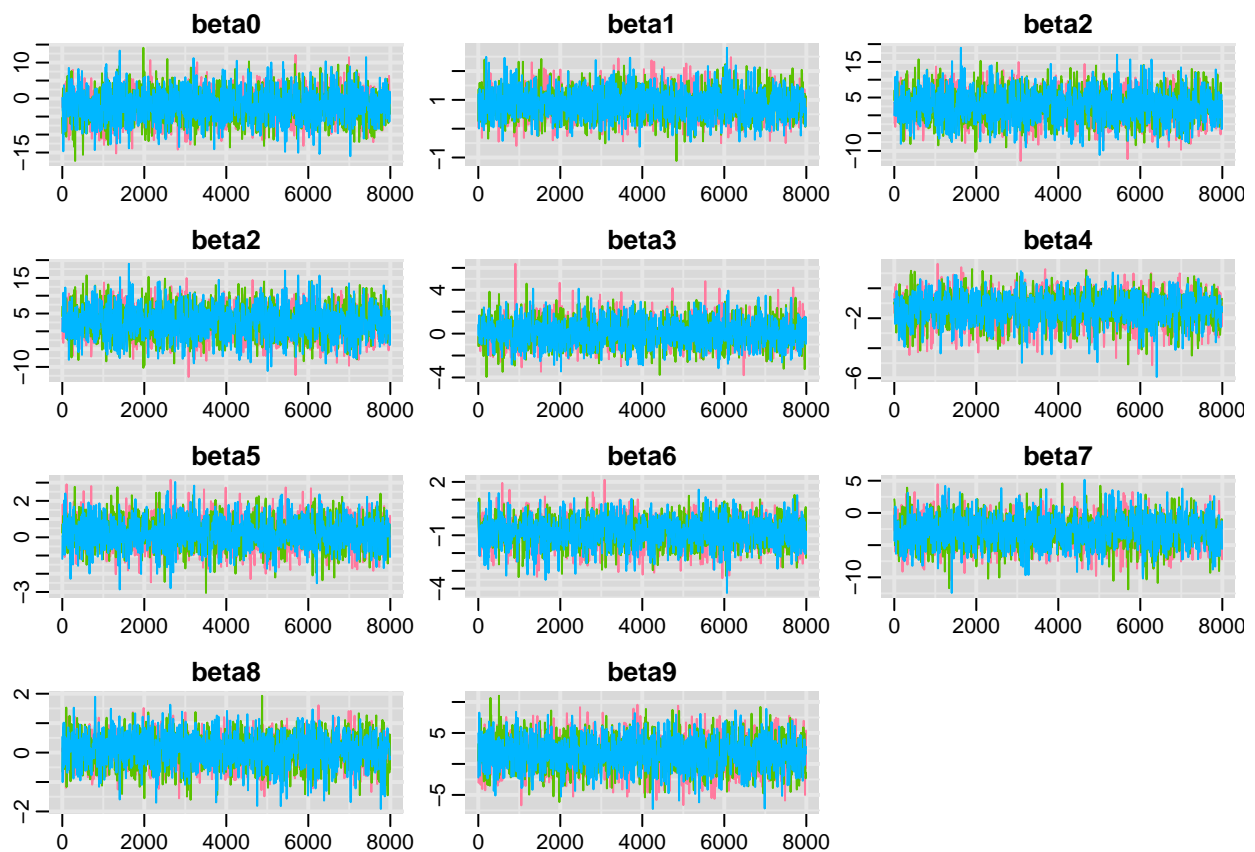
beta4 ~ dnorm(mu_0, sig_0)
beta5 ~ dnorm(mu_0, sig_0)
beta6 ~ dnorm(mu_0, sig_0)
beta7 ~ dnorm(mu_0, sig_0)
beta8 ~ dnorm(mu_0, sig_0)
beta9 ~ dnorm(mu_0, sig_0)
}",
file = "log_regr.txt")

```

For this specific model:

- three Markov chains are used;
- we set 2000 iterations;
- no strategy is applied for the exclusion of burn-in samples.

4.1 MCMC convergence analysis



4.2 Prediction

```
## [1] "Accuracy: 0.931034482758621"
```

5. Second model

```
cat("model {  
  for (i in 1:N) {  
    z[i] <- beta0 + beta1*Season[i] + beta2*Age[i]  
      + beta3*Disease[i] + beta4*Accident[i] + beta5*Surgery[i]  
      + beta6*Fever[i] + beta7*Alcohol[i] + beta8*Smoking[i]  
      + beta9*Sedentarity[i]  
    y[i] ~ dbern(pnorm(z[i], 0, 1))  
  }  
  
  # Priors  
  beta0 ~ dnorm(mu_0, sig_0)  
  beta1 ~ dnorm(mu_0, sig_0)  
  beta2 ~ dnorm(mu_0, sig_0)  
  beta3 ~ dnorm(mu_0, sig_0)  
  beta4 ~ dnorm(mu_0, sig_0)  
  beta5 ~ dnorm(mu_0, sig_0)  
  beta6 ~ dnorm(mu_0, sig_0)  
  beta7 ~ dnorm(mu_0, sig_0)  
  beta8 ~ dnorm(mu_0, sig_0)  
  beta9 ~ dnorm(mu_0, sig_0)  
}",  
file = "probit_regr.txt")
```