Bayesian Modelling Project

Alex Costa and Sara Pascali

2024-06-26

1 Introduction

The purpose of this project is to conduct a detailed Bayesian analysis on the potential threat posed by asteroids to Earth, using various characteristics of these asteroids. Specifically, we aim to explore how these features influence the probability of an asteroid being classified as hazardous or non-hazardous to our planet. The analyzed data are taken from the CNEOS website, NASA's Center for Near Earth Object Studies.

In the dataset the binary variable **"Hazardous"** indicates whether an asteroid is considered hazardous (1) or non-hazardous (0) based on criteria defined by the International Astronomical Union and it represents our response variable. About 15.5% of the 4687 asteroids in the dataset are Hazardous.

Through the use of some Bayesian models, we will identify the key factors that contribute to make an asteroid hazardous, predict the probability that an asteroid is hazardous, and use that probability to classify the asteroids as hazardous or non hazardous.

1.1 Description of the data set

In addition to the response variable, the dataset contains 15 predictors:

- **Absolute Magnitude**: A standardized measure of an asteroid's luminosity.
- **Estimated Diameter**: An estimate of the diameter of an asteroid expressed in kilometers.
- **Relative Velocity**: The speed at which an asteroid moves relative to Earth during its close approach, expressed in kilometers per second.
- **Miss Distance**: The minimum distance at which the asteroid will pass by Earth during a close approach, expressed in kilometers.
- **Orbit Uncertainty**: The uncertainty of the estimated orbital parameters: 1 = Low Uncertainty; 2 = Medium Uncertainty; 3 = High Uncertainty.
- **Minimum Orbit Intersection**: The minimum distance between the object's orbit and Earth's orbit, expressed in Astronomical Units (1 AU = 1,496e+8 Km).
- **Eccentricity**: The shape of the orbit, with values ranging from 0 (circular orbit) to 1 (highly elongated elliptical orbit).
- **Semi Major Axis**: One half of the longest diameter of the elliptical orbit, expressed in Astronomical Units. It essentially defines the size of the orbit.
- **Inclination**: The angle between the asteroid's orbital plane and the ecliptic plane (Earth's orbital plane), evaluated in degrees.
- **Orbital Period**: The time required for the asteroid to complete one full orbit around the Sun, expressed in years.

• **Perihelion Distance**: The distance from the asteroid's Perihelion (the closest point in the orbit to the Sun) to the Sun, expressed in Astronomical Units.

• **Perihelion Argument**: The angle between the ascending node and the Perihelion, expressed in degrees.

• **Aphelion Distance**: The distance from the asteroid's Aphelion (the farthest point in the orbit to the Sun) to the Sun, expressed in Astronomical Units.

• **Mean Anomaly**: The fraction of the orbit already completed by the asteroid since Perihelion, expressed in degrees.

• **Mean Motion**: The average angular speed at which the asteroid orbits the Sun, expressed in degrees per day.

1.2 Graphical viasualization of the variables

We split the 4687 observations of the dataset into two parts: 10% of the observations in the Test set and the remaining 90% in the Training set. We will use the Training set to fit the regression models and the Test set to assess their performances.

We start with an exploratory analysis of the data, presenting the most relevant plots below:

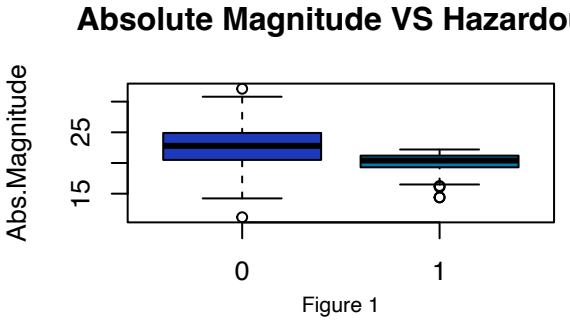

Figure 1

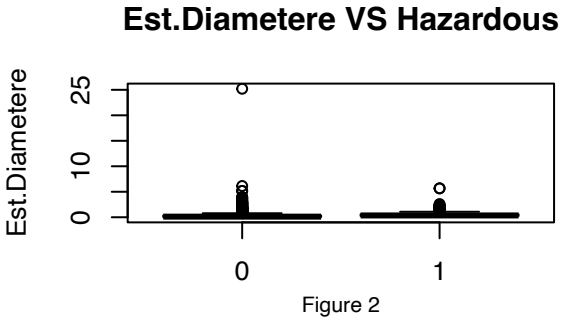

Figure 2

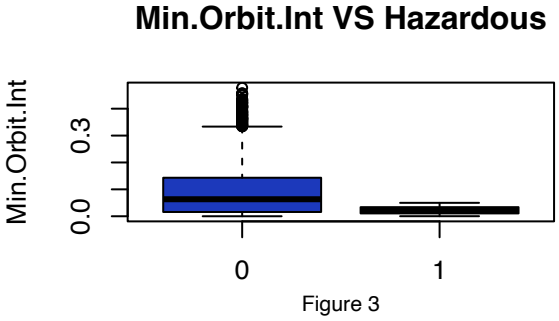

Figure 3

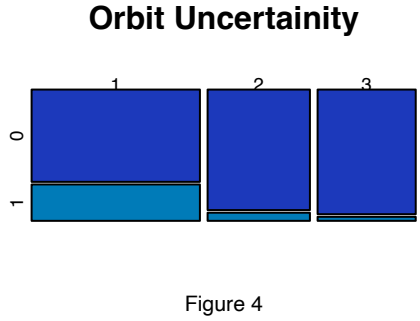

Figure 4

Figure 1 shows the absolute magnitude based on the hazardousness of the asteroids. It is observed that non-hazardous asteroids exhibit a greater variability in absolute magnitude compared to hazardous asteroids. Furthermore, it appears that hazardous asteroids tend to have a lower absolute magnitude, indicating they are more luminous.

From the Figure 2 we can note that Hazardous asteroids have slightly bigger quantiles for the diameter, but there is one non-hazardous asteroid which has a very big diameter, around 76 times bigger than the mean diameter of the sample.

Figure 3 shows that Hazardous asteroids have, in general, lower Minimum Orbit Intersection than non-hazardous asteroids. This makes totally sense because intuitively an asteroid who's orbit has a small distance from Earth's orbit is more likely to be dangerous. However, the variability for non-hazardous asteroids is higher, so that for small Minimum Orbit Intersections we also have non-hazardous asteroids.

The last plot, Figure 4, shows that 47% of the asteroids in the dataset have a Low orbit uncertainty, 28% have a Medium orbit uncertainty and 25% have an High orbit uncertainty. The majority of hazardous asteroids have a Low orbit uncertainty.

Before proceeding with the Bayesian analysis, we standardize the data by subtracting to each value of the predictors the sample mean of the corresponding variable. Markov Chain Monte Carlo (MCMC) sampling methods can converge more quickly and efficiently when the data are standardized. Additionally, standardization facilitates a simpler interpretation of the results.

2 GLM for binary response variables

A Generalized Linear Model (GLM) is a generalization of the ordinary linear regression model that allows the dependent variable to follow probability distributions different from the Normal.

In our case, the response is a binary variable, which means it can assume only two values: 0 if the asteroid is not dangerous or 1 if the asteroid is dangerous. Therefore, the dependent variable follows a *Bernoulli* distribution with π representing the probability that the asteroid is dangerous.

In a Generalized Linear Model, the response variable is linked to the linear predictor through a link function. In this analysis, we will apply and compare two of the most common link functions for binary responses: *logit* and *probit*.

2.1 Logistic regression

In logistic regression we assume a logit link function:

$$\pi_i = h(\boldsymbol{\beta}^T \boldsymbol{x}_i) = \frac{e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}}{1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}} \qquad \Leftrightarrow \qquad log(\frac{\pi_i}{1-\pi_i}) = \boldsymbol{\beta}^T \boldsymbol{x}_i$$

The sampling model is a Bernoulli distribution with π the probability of being hazardous, hence the *likelihood function* is:

$$p(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{i=1}^{n} p(y_i|\pi_i) = \prod_{i=1}^{n} h(\boldsymbol{\beta}^T \boldsymbol{x}_i)^{y_i} (1 - h(\boldsymbol{\beta}^T \boldsymbol{x}_i))^{1-y_i}$$

We assume an independent Normal *prior* for each of the beta parameters, so that the joint prior is the product of the marginal priors:

$$\beta_j \overset{ind}{\sim} N(\beta_{0j}, \sigma^2_{0j}) \quad \Rightarrow \quad p(\boldsymbol{\beta}) = \prod_{j=1}^{p} dN(\beta_j|\beta_{0j}, \sigma^2_{0j})$$

This prior is not conjugate to the model, therefore we will approximate the posterior distribution implementing a Metropolis-Hastings algorithm.

2.2 Probit regression

In probit regression we assume as link function the inverse of the cumulative distribution function of a standard Normal random variable:

$$\pi_i = h(\boldsymbol{\beta}^T \boldsymbol{x}_i) = \Phi(\boldsymbol{\beta}^T \boldsymbol{x}_i) \qquad \Leftrightarrow \qquad \Phi^{-1}(\pi_i) = \boldsymbol{\beta}^T \boldsymbol{x}_i$$

The likelihood and the prior distributions are the same as in logit regression. We can use again a Metropolis-Hastings algorithm, but in this case we can also adopt a latent variable interpretation, which allows to obtain the full conditional distributions in closed form, and therefore to implement a Gibbs Sampler.

So, by assuming that for every y_i there exists a latent variable z_i such that:

$$z_i|\boldsymbol{\beta} \overset{ind}{\sim} N(\boldsymbol{\beta}^T \boldsymbol{x}_i, 1) \qquad\qquad \boldsymbol{\beta} \sim N_p(\boldsymbol{\beta}_0, \boldsymbol{V}^{-1})$$

We obtain the following full conditional distributions for $\boldsymbol{z}$ and $\boldsymbol{\beta}$:

$$z_i|\boldsymbol{\beta}, \boldsymbol{y} \overset{ind}{\sim} tN(\boldsymbol{\beta}^T \boldsymbol{x}_i, 1, \theta_{y_i-1}, \theta_{y_i}) \qquad\qquad \boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{z} \sim N_p(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{V}}^{-1})$$

We can sample from these full conditionals and obtain a sequence of samples $(\boldsymbol{\beta}^{(1)}, ..., \boldsymbol{\beta}^{(S)})$ approximately from the posterior distribution $p(\boldsymbol{\beta}|\boldsymbol{y})$.

2.3 Prediction: Bayesian Model Averaging

Instead of performing a variable selection and use one selected model to make predictions, we will use the Bayesian Model Averaging method. Hence, the final prediction will be a sum of the predictions made with each of the possible models weighted by the corresponding posterior model probability.

We start by introducing a binary vector $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_p)^T$ such that:

$$\gamma_j = \begin{cases} 1 & \text{if } X_j \text{ is included in the model} \\ 0 & \text{if } X_j \text{ is not included in the model} \end{cases}$$

And we write the model as:

$$E(Y|x) = h(\gamma_1\beta_1 X_1 + ... + \gamma_p\beta_p X_p)$$

We assign the following priors to $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and w:

$$\beta_j|\gamma_j = 1 \overset{ind}{\sim} N(\beta_{0j}, \sigma^2_{0j}) \qquad\qquad \gamma_j \overset{iid}{\sim} Ber(w) \qquad\qquad w \sim Beta(a, b)$$

So that the joint prior on (β_j, γ_j) is the spike and slab prior:

$$p(\beta_j, \gamma_j) = (1 - w)\delta_0 + wdN(\beta_j|\beta_{0j}, \sigma^2_{0j})$$

Using Gibbs Sampling or Metropolis Hastings we obtain S draws from the posterior distribution of each β_j and γ_j. Now, given the values of the predictors for a new observation, $\boldsymbol{x}^*$, the BMA posterior predictive distribution for y* is given by :

$$p(y^*|\boldsymbol{y}) = \sum_{k=1}^{K} p(y^*|\boldsymbol{y}, M_k)p(M_k|\boldsymbol{y})$$

We can approximate it using the following algorithm. For s = 1,. . . ,S:

1) Compute $\eta^{(s)} = \gamma_1^{(s)}\beta_1^{(s)}x_1^* + ... + \gamma_p^{(s)}\beta_p^{(s)}x_p^*$

2) Compute $\mu^{(s)} = h(\eta^{(s)})$

3) Draw $y^{*(s)}$ from $p(y^*|\mu^{(s)})$

The output is a BMA sample from the predictive distribution of Y for subject i.

3 Applying the models to our data

3.1 Logit regression

We apply the logit regression model as described in chapter 2.1 to the asteroids data and, using a Metropolis-Hastings algorithm we approximate the posterior distribution of all the β_j. We set the following values for the prior hyperparameters of the β_j: $\beta_{0j} = 0 \quad \forall j, \qquad \sigma_{0j}^2 = 100 \quad \forall j.$

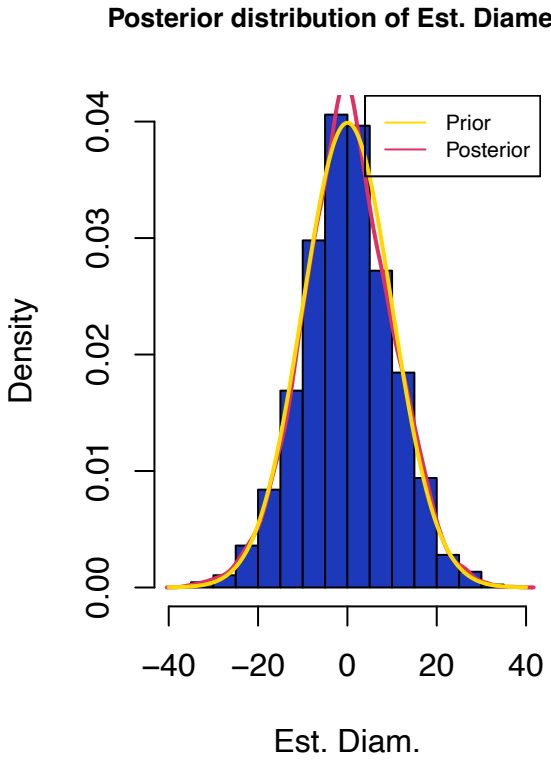

Figure 5

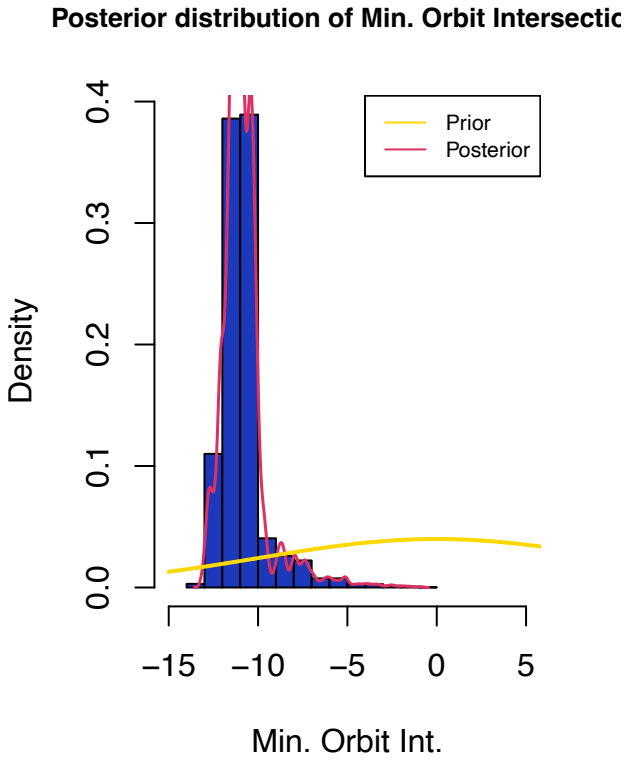

Figure 6

Figures 5 and 6 show the posterior distributions of Estimated Diameter and Minimum Orbit Intersection obtained with the logit regression, together with the corresponding priors. These plots provides a clear understanding of how much the data have contributed to modifying the initial beliefs about the parameters of interest. As we can see from Figure 5, both prior and posterior are very similar. This indicates that the observed data did not provide much new information beyond what was initially hypothesized. Conversely, in the Figure 6, the prior is very flat. This highlights that the observed data are highly informative and have significantly contributed to the formation of the posterior distribution.

To verify the goodness of the approximation provided by the Metropolis-Hastings algorithm we carry out a **diagnostics** using the following tools:

Trace plot: graphical representation of the sampled values $\theta^{(s)}$ across iteration $s = 1, \ldots, S$. For a good approximation, the chain should exhibit no trends and it should be concentrated within a region of high posterior probability, centered around the mode of $P(\theta|\boldsymbol{y})$.

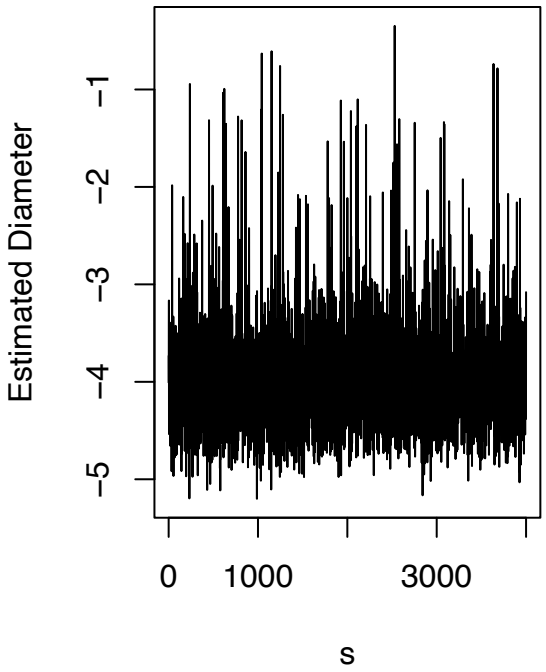

Figure 7

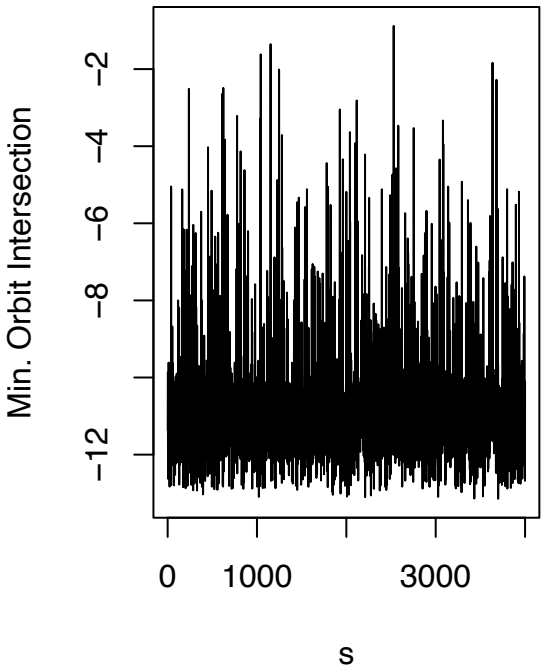

Figure 8

Figures 7 and 8 show the trace plots of the draws for the coefficients associated with the variables Estimated Diameter and Minimum Orbit Intersection. Both plots look nice, indeed there seem to be no trends and the values are quite concentrated in a region, even if there are some peaks of very high values. Furthermore, the burn-in period of 1000 draws is sufficient to reach convergence of the chain.

Auto Correlation Function: a measure of correlation in a Markov Chain. For a good approximation, the autocorrelation between various lags should decay rapidly towards zero. This indicates that successive observations are weakly correlated.

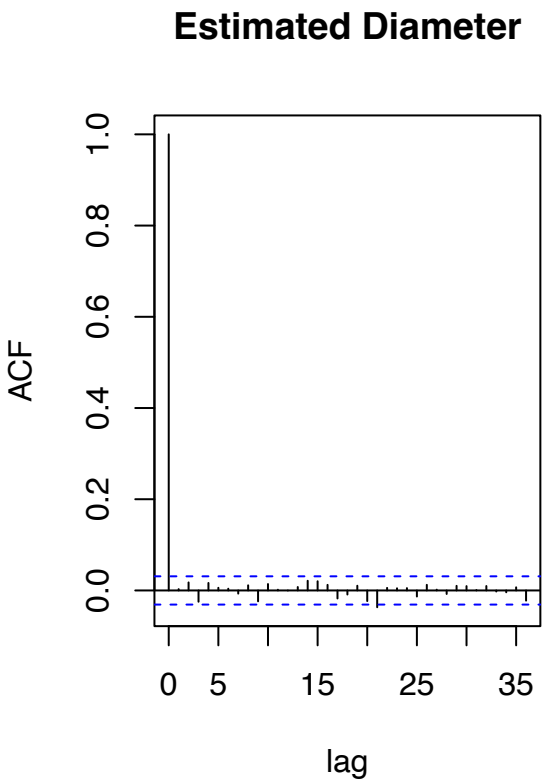

Figure 9

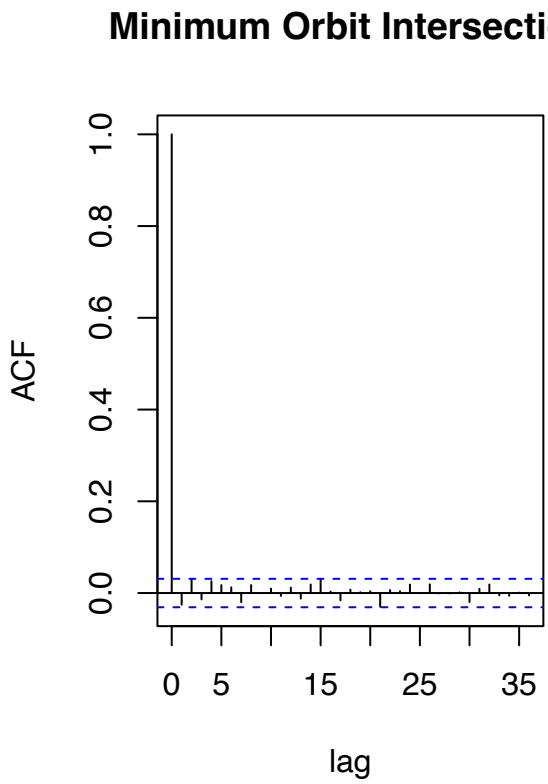

Figure 10

We can see from the ACF for Estimated Diameter (Figure 9) and Minimum Orbit Intersection (Figure 10) that the level of autocorrelation is very low for every lag, and this indicates that we don't need to thin the chain. It holds also for all the other coefficients.

Geweke test: a statistical test that compares the means of two subsets of the chain, one consisting of the first X% samples and the other consisting of the last Y% samples. If the chain converges, it is expected that these means are similar. Otherwise we should consider increasing the burn-in period. We choose X = Y = 10%.

beta[1]	beta[2]	beta[3]	beta[4]	beta[5]	beta[6]	beta[7]	beta[8]	beta[9]
-1.287	-1.5274	-1.1568	2.2741	0.9923	-0.4709	0.8798	-0.9866	-1.3076

beta[10]	beta[11]	beta[12]	beta[13]	beta[14]	beta[15]	beta[16]	beta[17]
-1.2661	1.589	-0.033	1.308	0.0611	-0.7433	1.7751	-1.2485

The output of the test are the values of the computed test statistic for each of the parameters in the model.The statistic used in the test is asymptotically Standard Normal, therefore at significance level 5% we reject the null hypothesis just for β_4, since its computed statistics is smaller than -1.96. Thus we can say that all the chains other that that have reached convergence and the burn-in period of 1000 draws is adequate.

Effective Sample Size: measures how much information is loss due to the correlation in the sequence. In other words, although our sequence may have a sample size of N, our effective sample size could be smaller or even bigger due to the correlation and redundancy between the samples.

beta[1]	beta[2]	beta[3]	beta[4]	beta[5]	beta[6]	beta[7]	beta[8]	beta[9]
3940.988	4000	4000	4000	4000	4258.263	4000	3972.359	4000

beta[10]	beta[11]	beta[12]	beta[13]	beta[14]	beta[15]	beta[16]	beta[17]
4000	3722.314	4000	4000	4000	4000	4000	3605.936

For most of the coefficients the Effective Sample Size is equal or almost equal to the sample size, 4000. Only one coefficient, β_{17}, has a significant decrease in the sample size due to autocorrelation.

3.2 Probit regression

We apply the probit regression model as described in chapter 2.2 to the asteroids data using a Metropolis-Hastings algorithm and not the latent variable interpretation. The values of the prior's hyperparameters are the same as in the previous chapter.

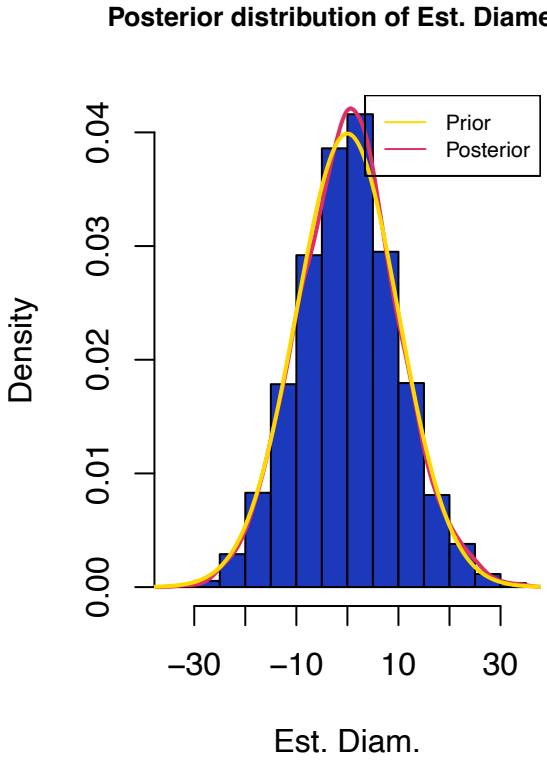

Figure 11

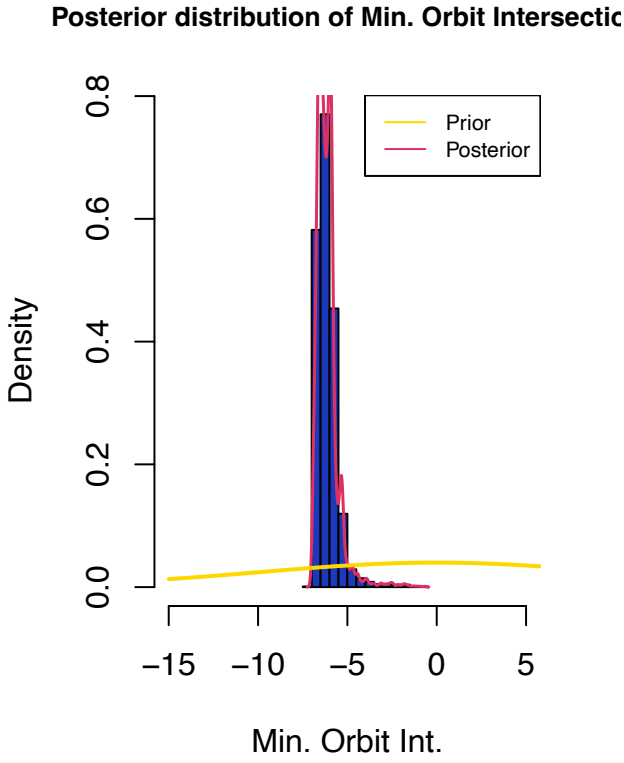

Figure 12

Figures 11 and 12 show the posterior distributions of Estimated Diameter and Minimum Orbit Intersection obtained with the probit regression, together with the corresponding priors. The posterior for the coefficient of Estimated Diamater is almost equal to the one obtained with the logit regression, while the one for Minimum Orbit Intersection has the same shape but it has an higher mean and it is even more concentrated around the mode.

We check again the **diagnostics**:

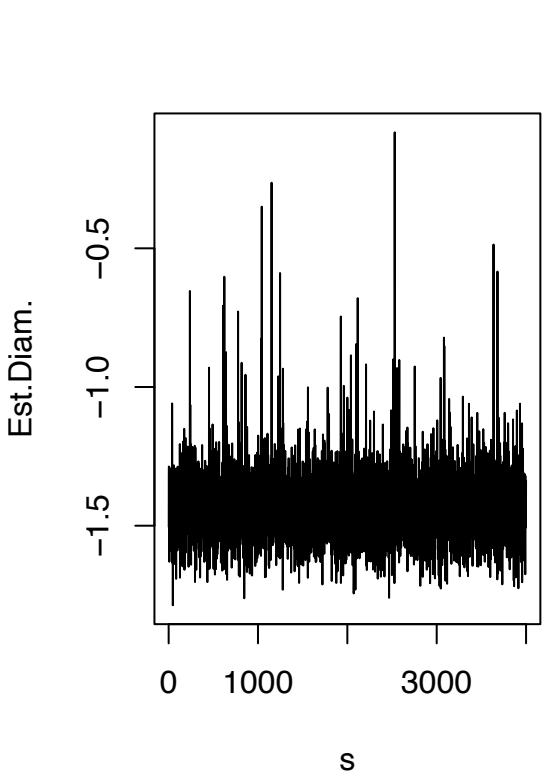

Figure 13

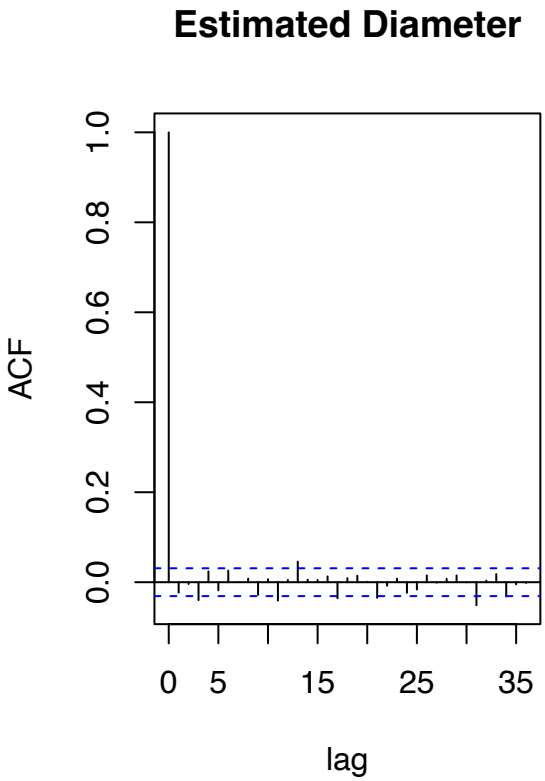

Figure 14

Figure 13 shows the trace plots of the draws for the coefficient of Estimated Diameter. Again there seem to be no trends, the values are quite concentrated in a region and the burn-in period is sufficient to reach convergence of the chain. Figure 14 shows the ACF for the same coefficient and we can say that there is a low level of autocorrelation at every lag, so we don't need to thin the chain.

Geweke test:

beta[1]	beta[2]	beta[3]	beta[4]	beta[5]	beta[6]	beta[7]	beta[8]	beta[9]
-1.287	-1.5274	-1.1568	2.2741	0.9923	-0.4709	0.8798	-0.9866	-1.3076

beta[10]	beta[11]	beta[12]	beta[13]	beta[14]	beta[15]	beta[16]	beta[17]
-1.2661	1.589	-0.033	1.308	0.0611	-0.7433	1.7751	-1.2485

Running the Geweke test, we reject again the null hypothesis just for β_4 at a 5% significance level.

Effective Sample Size:

beta[1]	beta[2]	beta[3]	beta[4]	beta[5]	beta[6]	beta[7]	beta[8]	beta[9]
3954.791	4319.731	4340.887	4000	3527.626	4000	4000	3860.62	3791.149

beta[10]	beta[11]	beta[12]	beta[13]	beta[14]	beta[15]	beta[16]	beta[17]
4000	4000	4000	4000	4000	4000	4000	4000

As a final diagnostic tool, we conducted the **Effective Sample Size** analysis and we obtained that the only coefficient that has a significant reduction in the sample size is β_5.

4 Prediction: comparison between Logit and Probit models

We now predict the probability of being hazardous for the asteroids in the test set using the BMA with both logit and probit models, we classify these asteroids based on a threshold, and we compare the results. We analyze below the **confusion matrix** for the predictions obtained with the two models. It is a tool used to visualize the model's predictions against the actual values, allowing us to see where the model is performing correctly and where it is making errors.

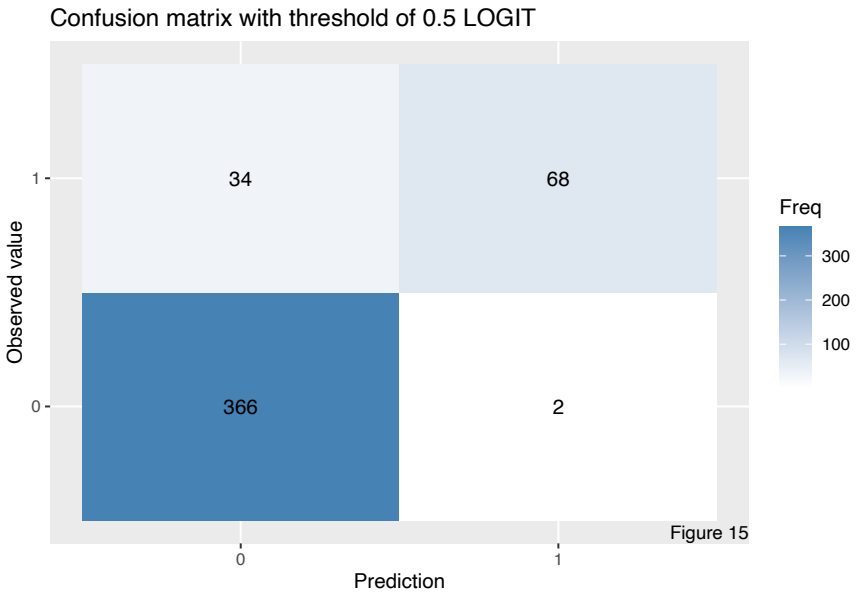

Figure 15 shows the confusion matrix for Logit model with the "standard" threshold of $\pi^* = 0.5$ for the probability required for an asteroid to be classified as hazardous. The performances of the logit model is pretty similar to the probit model. However, we decided to reduce the threshold to $\pi^* = 0.3$ in order to have less "False Negatives". Indeed having an asteroid classified as non-hazardous when in reality it is hazardous is much more dangerous than classifying it as hazardous when in reality it is not.

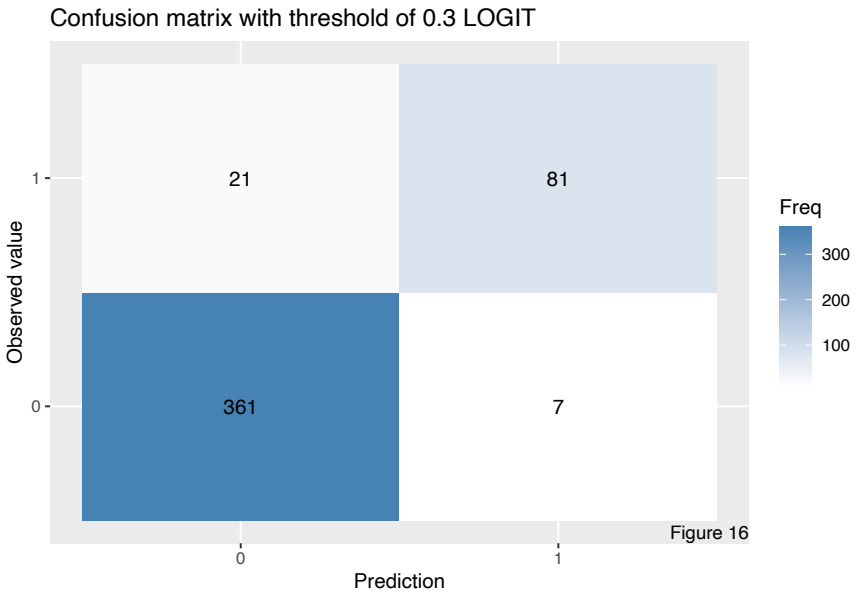

Figure 16 shows the confusion matrix for the Logit model with threshold of 0.3. Decreasing the threshold had the desired effect: the false negatives have decreased from 34 to 21 and the total number of wrong predictions has decreased from 36 to 28.

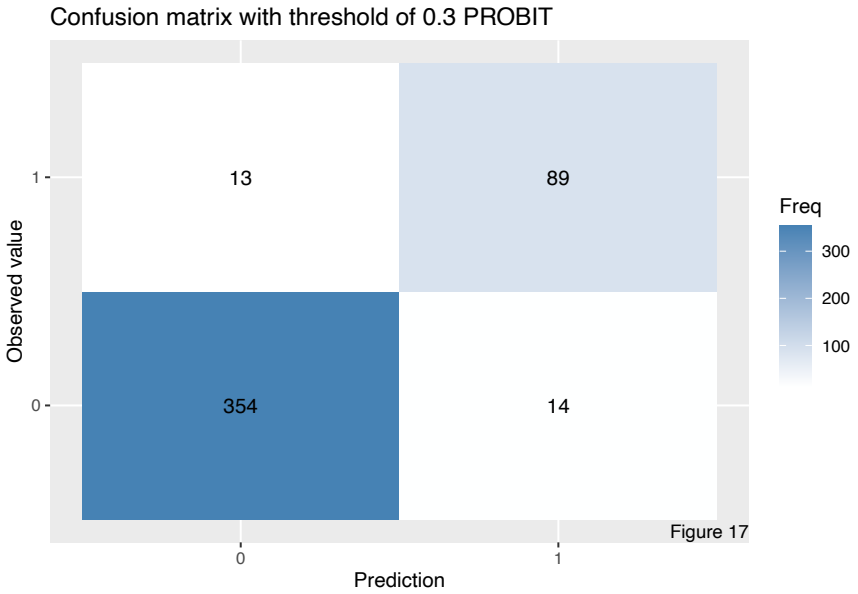

Figure 17 displays the confusion matrix for the Probit model using a threshold of 0.3. The Probit model shows a slight improvement over the Logit model by reducing false negatives by 8 and the total number of wrong predictions by 1. Furthermore, we can compute:

- The **accuracy**, that is the overall performance of correctly classified cases and it is equal to 94%.
- The **specificity**, that is the proportion of negatives that are correctly classified as such, and it is equal to 96%.
- The **sensitivity**, that is the proportion of positives that are correctly classified as such, and it is equal to 87%.

All these values indicate good performances.

5 Conclusion

In conclusion, this project has demonstrated that both the probit and probit models are effective tools for predicting the hazardousness of asteroids and are very similar in terms of prediction accuracy and algorithm convergence. Consequently, the choice between the two models can be guided by practical considerations or personal preference rather than significant differences in predictive accuracy.