Bayesian Data Analysis Mini Project Report

Group 6:

Aquila Kyne Sudiro / 2702212162
Caroline Ang / 2702208606
Laurel Evelina Widjaja / 2702213770

1. Introduction

The dataset chosen is “Mobile Device Usage and User Behavior Dataset”. The dataset consists of 11 variables, which are:

a. UserID
b. Device Model
c. Operating System
d. App Usage Time (min/day)
e. Screen On Time (hours/day)
f. Battery Drain (mAh/day)
g. Number of Apps Installed
h. Data Usage (MB/day)
i. Age
j. Gender
k. User Behavior Class → Target Variable

User ID	Device Mo...	Operating ...	App Usage...	Screen On...	Battery Dr...	Number of...	Data Usag...	Age	Gender	User Beha...
1	Google Pixel 5	Android	393	6.4	1872	67	1122	40	Male	4
2	OnePlus 9	Android	268	4.7	1331	42	944	47	Female	3
3	Xiaomi Mi 11	Android	154	4.0	761	32	322	42	Male	2
4	Google Pixel 5	Android	239	4.8	1676	56	871	20	Male	3
5	iPhone 12	iOS	187	4.3	1367	58	988	31	Female	3
6	Google Pixel 5	Android	99	2.0	940	35	564	31	Male	2
7	Samsung Galaxy S21	Android	350	7.3	1802	66	1054	21	Female	4
8	OnePlus 9	Android	543	11.4	2956	82	1702	31	Male	5
9	Samsung Galaxy S21	Android	340	7.7	2138	75	1053	42	Female	4
10	iPhone 12	iOS	424	6.6	1957	75	1301	42	Male	4

There are 700 rows of data for each variable and there are no null and duplicated data. UserID will be dropped since it has no effect on the classification model. Device Model, Operating System, and Gender will be encoded using label encoding otherwise the variable won’t be able to be used in the model. For User Behavior Class, originally it has 5 groups, however it’s changed to 0 (for data that was originally class 0, 1, 2) and 1 (for data that was originally class 3 and 4) instead of 0-4, in which 0 represents a user that doesn't use mobile device that often and 1 represents a user who often uses mobile device.

The objective of this case is to compare two Bayesian regression models for binary classification (Logistic regression with a logit link function and Probit Regression with a probit link function). Both models are compared based on posterior estimates, convergence diagnostics, and predictive accuracy.

2. Models

Two models used are Logistic Regression and Probit Regression because both of the models are used to classify a class, which suits the target of our dataset (variable Y), which is classifying whether a user falls in group 1 or 0. The predictor matrix (X) consists of 9 numerical variables. Both models use uninformative normal priors with mean 0 and very large variance 1000 ($\beta[j] \sim Normal(0,\ 0.001)$). These priors allow the posterior distributions of parameters to be primarily determined by the data rather than prior assumptions.

Both Logistic and Probit Regression model assumes a Bernoulli likelihood for Y: $Y[i] \sim Bernoulli(\pi[i])$. For Logistic Regression model, the probability of success ($\pi[i]$) for each observation is modeled using logit link function:

$logit(\pi[i]) = log(\frac{\pi[i]}{1-\pi[i]}) = \beta1 + Xi1\beta2 + ... + Xi9\beta10$, where:

- $Y[i]$: binary response for the i-th observation
- $\pi[i]$: probability of $Y[i] = 1$
- β : regression coefficient for predictors X

While for Probit Regression model, the probability of success ($\pi[i]$) for each observation is modeled using probit link function:

$probit(\pi[i]) = \Phi(\pi[i]) = \beta1 + Xi1\beta2 + ... + Xi9\beta10$, where:

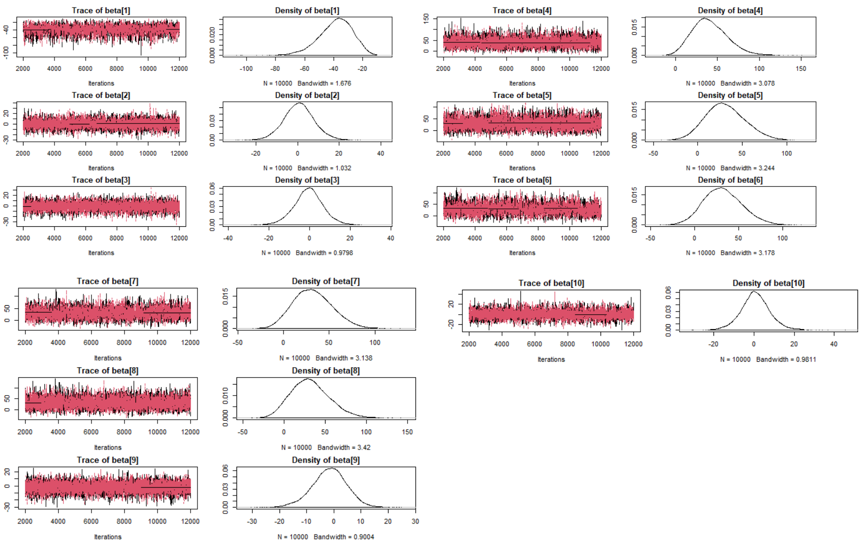
- Φ^{-1} : the inverse of the cumulative distribution function (CDF) of the standard normal distribution
- $\pi[i]$: probability of unobserved variable exceeding a certain threshold

3. Algorithm

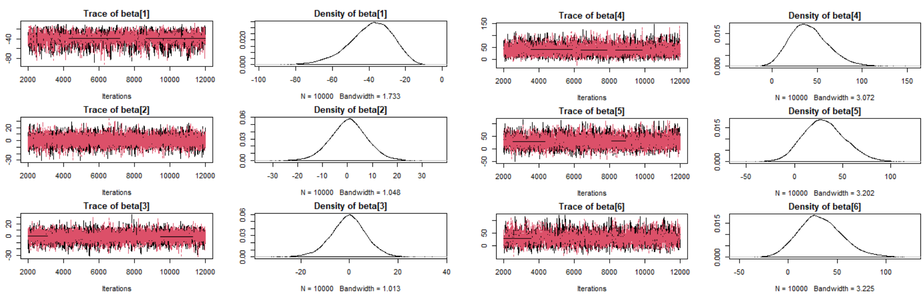
The models are implemented using Markov Chain Monte Carlo (MCMC) methods via JAGS. Here are the settings of the MCMC algorithm:

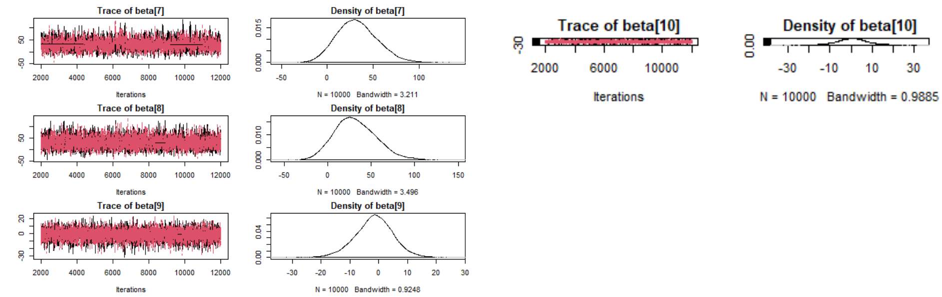
- Chains: 2 independent Markov chains
- Burn-in: 1000 iterations discarded to allow the chains to stabilize
- Sampling: 10000 iterations are collected after burn-in

Below are the trace plots that visualize the Logistic regression sampled parameter ($\beta0, \beta1, ..., \beta10$) values across iterations for each chain. The Logistic Regression model can be considered as convergent as it shows a stable and consistent horizontal band with no obvious trends, indicating stationarity.



Below are the trace plots that visualize the Probit Regression sampled parameter ($\beta0, \beta1, ..., \beta10$) values across iterations for each chain. Because it also shows a stable band, the Probit Regression model can be considered as convergent too.





Model Parameter		ESS	Gelman Diagnostic	Geweke Diagnostic Chain 1	Geweke Diagnostic Chain 2
Logistic Regression	beta[1]	1774.277	1.00	-2.1973	0.4882
	beta[2]	5697.880	1.00	-0.3836	-0.5526
	beta[3]	6831.159	1.00	0.4530	0.8612
	beta[4]	3566.528	1.00	1.0590	-0.9306
	beta[5]	3532.041	1.00	-0.1160	0.4918
	beta[6]	2257.077	1.00	1.8747	-0.1880
	beta[7]	2767.933	1.00	0.2408	-1.0412
	beta[8]	4309.682	1.00	0.8437	0.6850
	beta[9]	7208.075	1.00	-0.3980	1.5709
	beta[10]	7130.600	1.00	-0.6063	1.0958
Probit Regression	beta[1]	1390.544	1.00	0.3799	-0.09701
	beta[2]	5520.685	1.00	0.5517	-0.47361
	beta[3]	5927.287	1.00	1.0607	-0.52321
	beta[4]	3251.506	1.00	0.6842	0.24176
	beta[5]	3241.251	1.00	-0.9312	-0.02787
	beta[6]	1985.420	1.00	-1.7455	0.34160
	beta[7]	2288.305	1.00	0.9535	-0.86342
	beta[8]	3912.066	1.00	-0.2693	0.17274
	beta[9]	6908.822	1.00	-0.4038	1.80328
	beta[10]	7194.657	1.00	-0.3351	1.74006

Based on the table above, both models have sufficiently high ESS, indicating that the MCMC chains produce enough effective samples for reliable posterior inference. Although the probit model generally has lower ESS compared to Logistic Regression, the values are still acceptable. For all parameters in both models have achieved convergence across all chains, as indicated by the Gelman-Rubin diagnostic = 1.00 (less than 1.1). In Logistic model chain 1, there are some parameters with Geweke values outside the [−2,2] range, such as beta[1] and beta[6]. But, chain 2 shows mostly stable values within [-2,2]. While for probit model chain 1 and chain 2 have values close to 0 for most parameters. However, beta[6] in chain 1 and beta[9] in chain 2 approach the boundaries of significance. Overall, the Geweke diagnostics show that the chains for both models are largely stationary. The minor deviations for a few parameters are not critical since the Gelman-Rubin diagnostic and ESS values confirm good convergence and efficiency.

4. Results

Based on DIC value, probit model (0.3099) has a lower mean deviance compared to the logistic model (0.5305). It indicates that probit model fits the data slightly better than logistic model. Logistic Regression doesn't have a penalty or NaN, usually it's caused by overfitting, meanwhile Probit Regression has both penalty and penalized deviance which are 12.03 and 12.34 respectively. Which means, probit model balances both fit and complexity better than the logistic model. Furthermore, based on the WAIC value, Probit Regression (0.56) has a lower number than Logistic Regression (0.95). Lower WAIC number usually suggests a better fitting model. So, it can be concluded that Probit Regression shows a better performance compared to Logistic Regression and it is a more reliable choice.

Next, some calculations are done from probit sample and it could be seen that both beta[1] and beta[4] didn't include zero, so both beta are considered significant. However, from the data shown below, it could be seen that beta[1] has a negative number, meanwhile beta[4] has positive number.

	Mean	SD	Naive SE	Time-series SE
beta[1]	-38.6484	11.525	0.08149	0.27437
beta[2]	0.6019	7.379	0.05218	0.09786
beta[3]	-0.5345	7.181	0.05077	0.08688
beta[4]	40.9035	21.077	0.14904	0.35531
beta[5]	30.9555	22.178	0.15682	0.37359
beta[6]	32.1660	21.728	0.15364	0.45938
beta[7]	32.0009	21.455	0.15171	0.40777
beta[8]	31.8595	23.482	0.16604	0.35783
beta[9]	-1.6630	6.485	0.04585	0.07639
beta[10]	0.7726	7.373	0.05213	0.08734

	2.5%	25%	50%	75%	97.5%
beta[1]	-63.776	-46.186	-37.7458	-30.471	-18.91
beta[2]	-14.283	-4.197	0.4558	5.180	15.82
beta[3]	-15.561	-5.095	-0.4178	4.016	13.80
beta[4]	4.927	26.065	39.5332	54.991	86.75
beta[5]	-8.815	16.033	30.2345	45.862	78.22
beta[6]	-6.942	17.078	31.6841	47.467	77.34
beta[7]	-7.756	16.292	30.4370	45.829	77.54
beta[8]	-10.566	14.689	30.0313	47.150	81.24
beta[9]	-15.484	-5.663	-1.3804	2.680	11.48
beta[10]	-14.037	-3.746	0.8316	5.434	16.59

It can be concluded that beta[1] or Device Model has a strong negative relationship with the outcome, which means an increase in Device Model value leads to a decrease in the probability of the outcome being 1. On the other hand, beta[4] or Screen on Time (hours/day) has a strong positive relationship with the outcome. So, an increase in Screen on Time value leads to an increase in the probability of the outcome being 1.

From the posterior check performed on this model, p-value of both sumY and meanY are the same, 0.05735. However, if we slightly change the model and make it as mean(D[, j] >= D0[j]) instead of '>', the p-value would be 0.9636. This shows that the model isn't the best option, because both p-values are located on the tail side, which means the model might be underfitting or overfitting.

5. Conclusion

Overall, from the two models chosen, Probit and Logistic Regression, it could be considered that probit is a slightly better model compared to logistic. However, there are still a lot of improvements required for this model in order for it to be accurate and trustable.