Bayesian Data Analysis Group Project 1

Group E (Nilufar Ibrahimli, Naila Ismayilova, Xiao Wang)

Data reprocessing and summary statistics:

Firstly, we started the dataset for further analysing and modelling. We created a subset of dataset related to loan approval, financial status, demographic factors, and credit history. The selected columns are approve, hrat, obrat, loanprc, unem, dep, sch, cosign, pubrec, married, white, and male. Then, we converted all variables into numeric format to ensure consistency. Non-convertible data types were set to missing values ("Nan"). Then, those rows containing missing values were eliminated from the dataset to ensure data integrity. Summary statistics were obtained including measures such as mean, standard deviation, minimum, maximum, and quartiles.

Figure 1.

	approve	hrat	obrat	loanprc	unem	dep
count	1971.000000	1971.000000	1971.000000	1971.000000	1971.000000	1971.000000
mean	0.876205	24.800081	32.389797	0.770431	3.888534	0.771689
std	0.329431	7.130267	8.276594	0.189467	2.171818	1.105423
min	0.000000	1.000000	0.000000	0.021053	1.800000	0.000000
25%	1.000000	21.000000	28.000000	0.700000	3.100000	0.000000
50%	1.000000	25.800000	33.000000	0.800000	3.200000	0.000000
75%	1.000000	29.000000	37.000000	0.898906	3.900000	1.000000
max	1.000000	72.000000	95.000000	2.571429	10.600000	8.000000

	sch	cosign	pubrec	married	white	male
count	1971.000000	1971.000000	1971.000000	1971.000000	1971.000000	1971.000000
mean	0.770167	0.028919	0.068493	0.659564	0.846271	0.813293
std	0.420832	0.167622	0.252654	0.473976	0.360780	0.389775
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	0.000000	0.000000	0.000000	1.000000	1.000000
50%	1.000000	0.000000	0.000000	1.000000	1.000000	1.000000
75%	1.000000	0.000000	0.000000	1.000000	1.000000	1.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

According to the results, the mean of the loan approval rate 0.87 (87.6%) indicates that most loan applications were approved. The minimum value is 0 which indicates rejections, and the maximum value is 1 which indicates the approval. The standard deviation of 0.33% suggests that there is some variability, but approval dominates the dataset.
The mean housing expense ratio is 24.8, while standard deviation is 7.13, which suggests a moderate spread. The minimum value is 1 which is unusually low, while the maximum value

reaches 72, indicating that some applicants allocate most part of their income to housing expenses.
The mean other obligations ratio is 32.29, with a slightly larger spread of 8.29. The maximum value is 95, which suggests that some applicants have a high debt burden.
The average loan-to-price ratio is 0.77 (77%), which means that, on average, applicants finance 77% of the value of their property with a loan. The minimum value is 0.02 (2%), while the maximum reaches 2.57, which suggests some cases where loans exceed the value of the property.
The average unemployment rate in the data set is 3.89, with a range of 1.8 to 10.6.
The relatively low standard deviation (2.17) suggests that most applicants are in regions with stable unemployment rates.
The average number of dependents is 0.77, but the range varies greatly between 0 and 8, suggesting that some applicants have large households. At least half of the applicants have no dependents (the median is 0).
Most applicants are married (binary with the mean of 66% and median of 1 which indicates being married), and have completed schooling (binary with the mean of 77% and all percentiles of 1).
Co-signers and public credit issues are rare in the dataset as median and 75th percentile is 0 for both of them.
The data set is not balanced in terms of race (0.85 white) and gender (0.81 male), which could have consequences for the fairness of loan approvals.

Bayesian probit model via gibbs sampling and interpretation of results:

We estimated the Bayesian probit model with gibbs sampling. The binary response variable (approve) was denoted as y. The other variables from the cleaned dataset were used as predictor variables, denoted as x, and an intercept term was added. 15,000 samples were drawn with the first 5,000 being treated as burn-in to allow convergence. Truncated normal distribution was used to sample latent variable y_star. The latent variable y_star is an unobserved continuous variable in the probit model. This latent variable is assumed to be generated from a normal distribution, and the binary outcome y is derived from y_star through a threshold process. The posterior distribution of the regression coefficients (beta) was updated iteratively, by adding a Gaussian prior with a mean of zero and a high variance covariance matrix (diagonal matrix with variance 1000), which represents a weakly informative prior. The average of the subsequent samples for each coefficient was calculated as the Bayesian estimate. The 95% credible intervals (CIs) were derived from the 2.5th and 97.5th percentiles of the posterior samples, providing a quantification of the uncertainty for each coefficient. Below you can find obtained results:

Figure 2.

	Variable	Mean	2.5% CI	97.5% CI
0	Intercept	2.396405	1.802504	2.967996
1	hrat	0.012020	-0.001273	0.025515
2	obrat	-0.031225	-0.043098	-0.019707
3	loanprc	-1.011999	-1.478952	-0.546403
4	unem	-0.035884	-0.069278	-0.002366
5	dep	-0.046109	-0.118807	0.027292
6	sch	0.034944	-0.154702	0.219962
7	cosign	0.056100	-0.376633	0.539319
8	pubrec	-0.988777	-1.229171	-0.750965
9	married	0.262254	0.077856	0.438948
10	white	0.592166	0.411237	0.773911
11	male	-0.060538	-0.273262	0.153992

An intercept is positive and significant, which suggests that when all other variables are in their baseline (typically 0), the loan rate approval probability is high. This matches with the high approval rate (87.6) in the dataset.

Housing expense ratio has a small positive coefficient meaning that higher housing expenses might slightly increase the loan approval rate. However, the coefficient is statistically insignificant as the credible interval (CI: -0.001 to 0.026) includes 0. Lenders will already be able to take housing costs into account in other financial indicators, reducing their unique impact.

Other debt obligations coefficient is negative and significant (CI: -0.004 to -0.019), which indicates that higher non-housing liabilities are less likely to be approved. High existing liabilities might be viewed as a risk factor by lenders.

Loan-to-price coefficient is negative and significant, with large magnitude. A higher ratio reduces loan approval likelihood. This indicates that lenders prefer applicants who can contribute more liquidity.

Unemployment rate has a statistically significant but negative coefficient, meaning that higher unemployment rates reduce the approval probability. Lenders likely prefer stable income when giving off loans.

Number of dependents coefficient is not statistically significant. It might be evident that having more dependents might reduce the income of applicants whilst decreasing the probability of loan approval. However, this factor does not influence the outcome in this dataset.

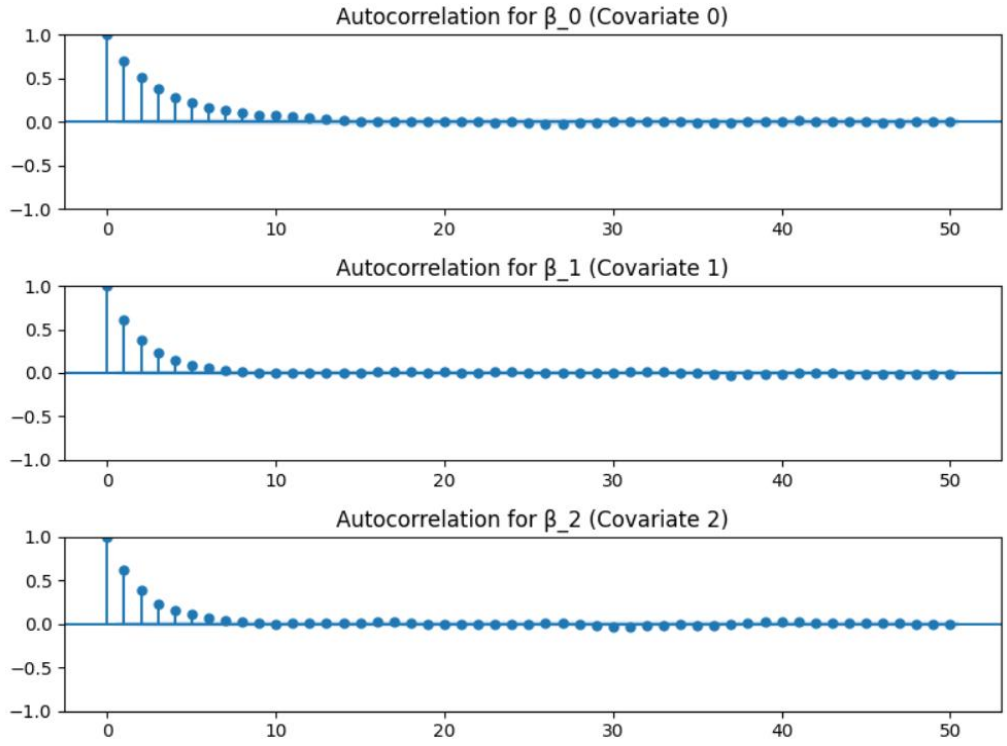
Schooling and co-signer coefficients are statistically insignificant. This might indicate that education alone is not a decisive factor for a lender, because income and credit history probably has higher importance. Similarly, having a co-signer might have increased the loan

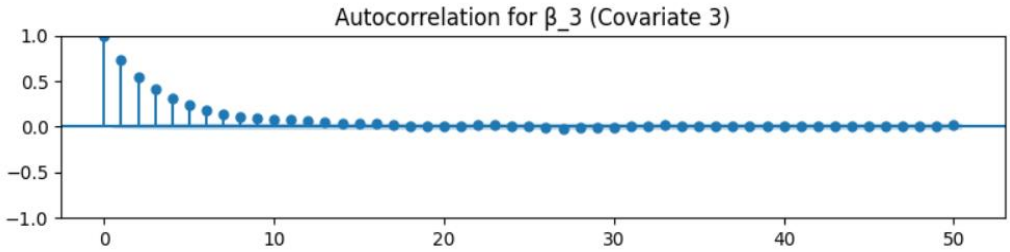
approval in terms of sharing the debt burden, however this variable is not a key determinant in this dataset.
Public credit records (e.g. bankruptcies) are strongly negative and highly significant, indicating that applicants with higher scores are much less likely to get an approval. This factor aligns with lending practices as lenders do not prefer a history of financial mismanagement.
The married applicants are more likely to obtain loan approval with a positive and statistically significant coefficient. This could be due to perceived financial stability, dual-income households, or conservative lending practices.
Applicants who belong to the white race have a higher probability of getting an approval, which may indicate racial disparities in the lending practice.
Finally, the loan decision does not differ between male and female applicants, as the male coefficient is statistically insignificant.
In conclusion, financial risk factors matter the most for obtaining an approval as high other debt obligations and a high loan-to-price ratio strongly reduce the approval chances. Additionally, a public credit record is the most detrimental factor for approval. Marital status and race also show significant associations as lenders prefer applicants who are married and white.

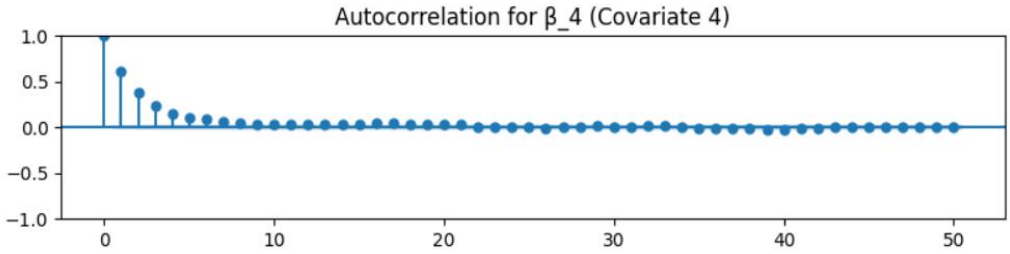
Autocorrelation analysis of gibbs sampling chain:

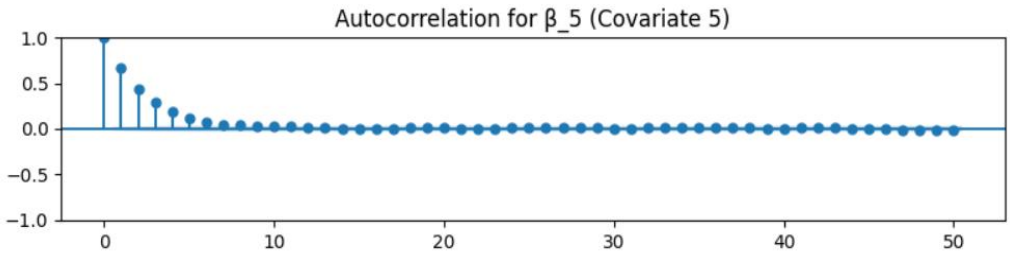
We visualized Autocorrelation Function (ACF) for each coefficient obtained from the gibbs sampling process. Autocorrelation in Markov Chain Monte Carlo (MCMC) techniques indicates how strongly each sample depends on the previous sample.

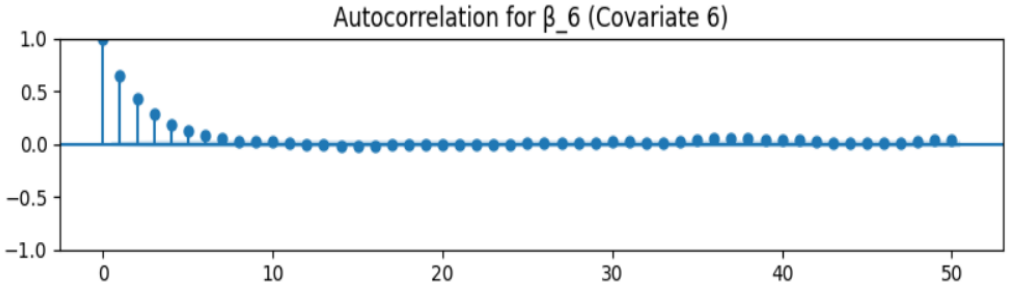
Figure 3.

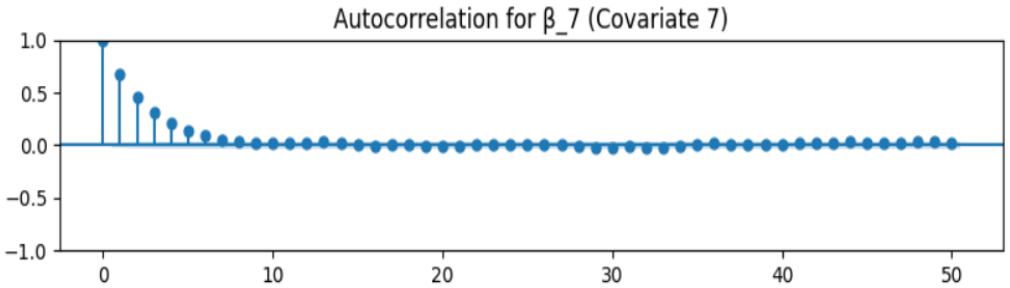


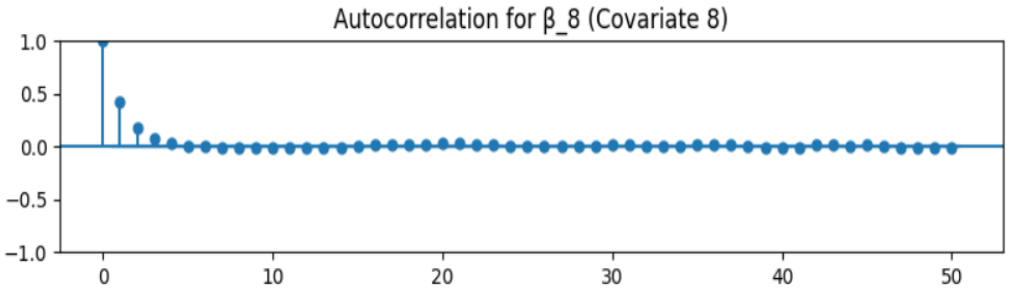

Autocorrelation for β_3 (Covariate 3)
1.0
0.5
0.0
−0.5
−1.0
0
10
20
30
40
50

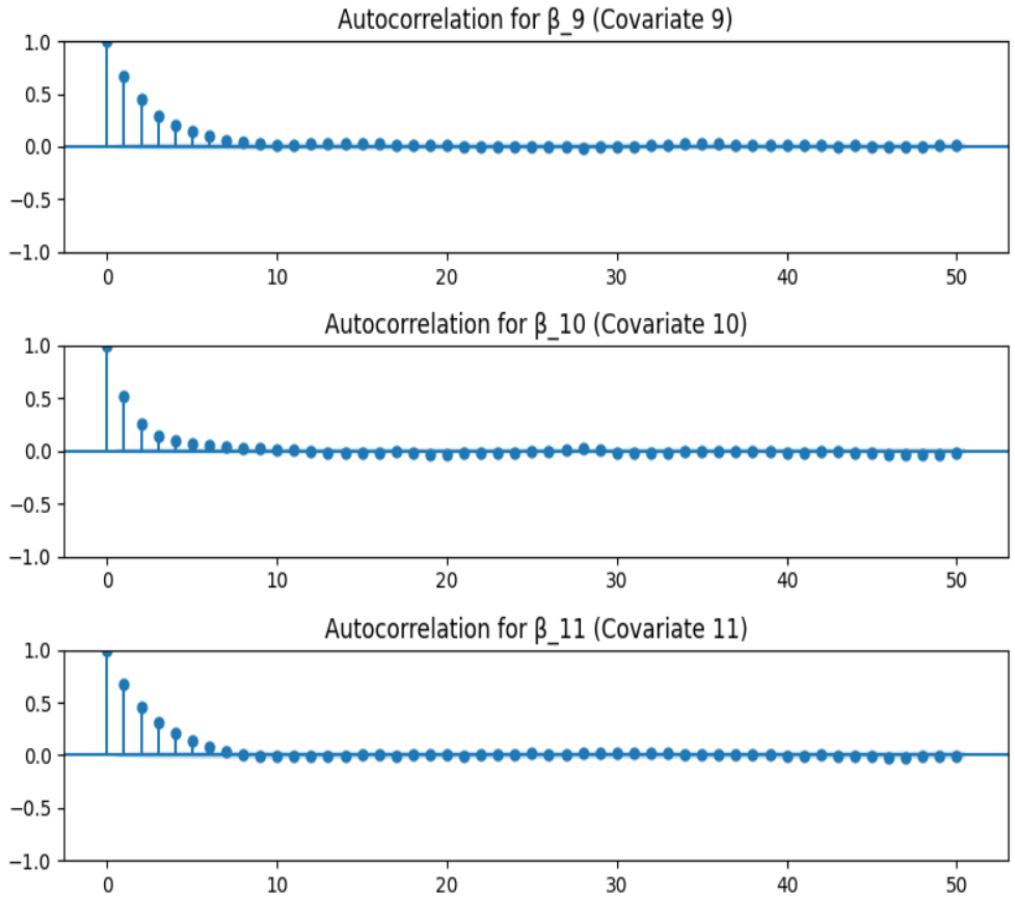

Autocorrelation for β_4 (Covariate 4)
1.0
0.5
0.0
−0.5
−1.0
0
10
20
30
40
50


Autocorrelation for β_5 (Covariate 5)
1.0
0.5
0.0
−0.5
−1.0
0
10
20
30
40
50


Autocorrelation for β_6 (Covariate 6)
1.0
0.5
0.0
−0.5
−1.0
0
10
20
30
40
50


Autocorrelation for β_7 (Covariate 7)
1.0
0.5
0.0
−0.5
−1.0
0
10
20
30
40
50


Autocorrelation for β_8 (Covariate 8)
1.0
0.5
0.0
−0.5
−1.0
0
10
20
30
40
50

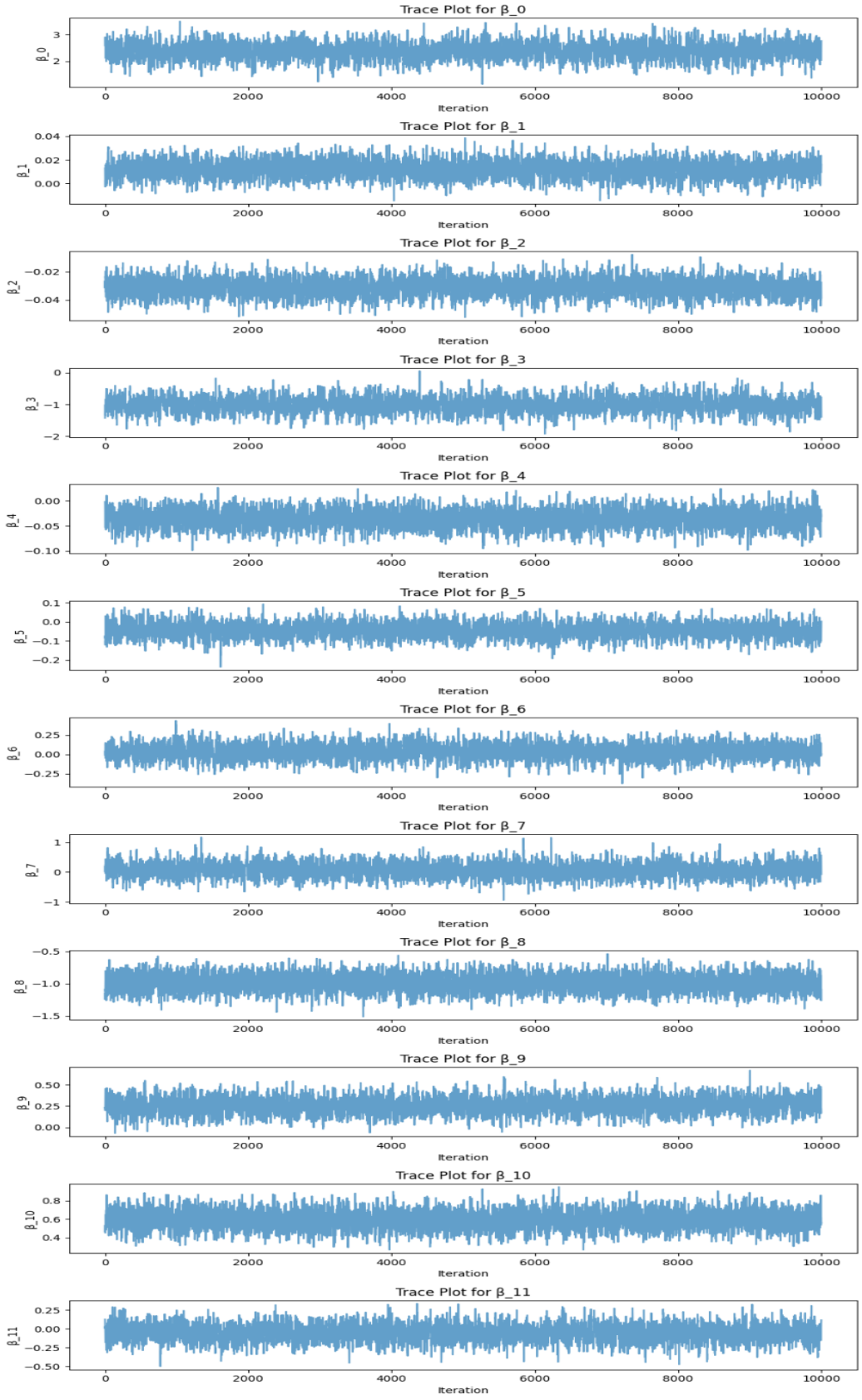


Based on the visual results, we can conclude that all coefficients show a strong positive correlation at initial lags. However, the correlation gradually decreases as the number of lags increases. The autocorrelation drops off within 10 lags, which suggests reasonable mixing of the Markov chains. If the autocorrelation persisted for longer lags, it would mean slow mixing and inefficient exploration of the posterior distribution. By the time the chain reaches lag 10-20, the autocorrelation is close to 0, which implies that the samples become nearly independent after a certain number of iterations. Some coefficents (e.g., β_0 (Intercept), β_3 (loanprc), β_8 (pubrec)) depict slightly slower geometrical decay than others, indicating that they may require a large number of iterations for efficient mixing.
In summary, due to the quick decline of autocorrelation, the sampler is efficient and likely providing a good representation of the posterior distribution, and the burn-in period of 5000 samples seems appropriate.

Trace plots analysis:

We plotted trace plots to visualize the sampled values of each regression coefficient across iterations of the Gibbs sampling. These plots help to determine the convergence, mixing, and stationarity of the MCMC process.

Figure 4.


Trace Plot for β_0
β_0
Iteration
Trace Plot for β_1
β_1
Iteration
Trace Plot for β_2
β_2
Iteration
Trace Plot for β_3
β_3
Iteration
Trace Plot for β_4
β_4
Iteration
Trace Plot for β_5
β_5
Iteration
Trace Plot for β_6
β_6
Iteration
Trace Plot for β_7
β_7
Iteration
Trace Plot for β_8
β_8
Iteration
Trace Plot for β_9
β_9
Iteration
Trace Plot for β_10
β_10
Iteration
Trace Plot for β_11
β_11
Iteration

Based on the results, we can see that the chains for all the coefficients fluctuate around a central value with no visible trend, indicating that the Markov chain has probably converged to its stationary distribution. The chains move quickly through the parameter space without getting stuck, which suggests good mixing.

Gelman-Rubin R-hat testing:

The Gelman-Rubin test is used when you run multiple chains as given in our case to assess the convergence of the MCMC process. This test compares the variance between chains to the variance within each chain. We reshaped original beta samples into 4 pseudo-chains each with 5,000 iterations and 6 parameters being estimated. In addition to R-hat, we assessed the efficiency of the MCMC process using the Effective Sample Size (ESS), which measures how many effectively independent samples are available. A higher ESS indicates fewer autocorrelations between samples and suggests that the samples are more representative of the posterior distribution.

Figure 5.

```
Data variables:
    beta     (param) float64 1.0 1.001 1.0 1.0 1.0 1.001
```

If R_hat is approximately 1, it indicates good convergence. If R_hat is much larger than 1 (>1.1), it indicates poor convergence. According to the obtained results, R_hat values for all pseudo-chains are approximately 1, which implies that chains have converged well to the posterior distribution. There are no significant differences between chains, which is an indicator of good mixing.

Figure 6.

```
Data variables:
    beta     (param) float64 1.754e+04 4.529e+03 ... 1.808e+04 3.905e+03
```

ESS measures independent information in the chain. High ESS exhibits more reliable estimates and more independent samples. Low ESS exhibits the high autocorrelation, suggesting longer chains. Our ESS values are in a range from 3,905 to 18,080. All ESS values are well above 1,000 with the highest value of 18,000, which indicates strong independent information from the samples.

Geweke diagnostics:

We did an additional test for convergence using Geweke diagnostics. It checks whether the mean of the first part of the MCMC process is equal to the mean of the last part. It calculates a Z-score, where values close to 0 suggest good convergence, while values beyond ±1.96 may indicate potential issues. A large absolute value of the Geweke statistics indicates that the chain has not converged yet.
We checked the mean and variance of the first 10% and the last 50% of the sampled values for each regression coefficient using Geweke diagnostic.

Figure 7.

	Variable	Geweke Z-score
0	[Intercept, hrat]	1.586970
1	[Intercept, obrat]	1.058825
2	[Intercept, loanprc]	-2.519440
3	[Intercept, unem]	0.135787
4	[Intercept, dep]	-1.589626
5	[Intercept, sch]	4.335037
6	[Intercept, cosign]	-2.347054
7	[Intercept, pubrec]	-0.505305
8	[Intercept, married]	-0.260149
9	[Intercept, white]	-0.321967
10	[Intercept, male]	0.614592
11	[Intercept]	-0.817951

Based on the results, Z- scores of the most variables are within acceptable range, indicating that their chains have converged. Loan-to-price ratio (-2.519), schooling (4.335), and cosigner (-2.347) fall out outside of the threshold ±1.96, showing potential non-convergence. However, this is not necessarily an issue, because the autocorrelation (as observed before) for these parameters is low, meaning that the sampling explores the parameter spaces efficiently. In addition to this, trace plots showed us well-mixed chains that stabilized over a period, which indicates good convergence. Since Geweke diagnostic relies on first and last sampled values, small deviations can sometimes occur even if the chain has converged properly.

Other variables, such as unemployment (0.135), marital status (-0.26), and race (-0.321) have their Z-scores close to 0, which implies stable convergence. Lastly, an intercept (-0.818) is also within acceptable range, showing no convergence problem.