# Bayesian Inference

Biljana Vitanova
*MLDS1 2024/25 , FRI, UL*
*bv7063@student.uni-lj.si*

## I. Introduction

In this project, we used both MCMC and Laplace approximation to estimate the posterior distribution of a Poisson regression model for predicting goals scored. We apply different diagnostics to ensure convergence of the chains. We analyze the estimated coefficients. Furthermore, we make point predictions based on different loss functions.

## II. Model Implementation

### A. Methodology

In the first part, we fit a Poisson GLM with a log link function. The model is defined as

$$\mu = \beta_0 + \beta_1 x_1 + \ldots, \quad \lambda = \exp(\mu),$$

where $\lambda$ is the expected number of goals.We standardize the explanatory variables to make the coefficients reliable to interpret. To estimate the parameters, we use Bayesian inference with MCMC. The MCMC is run with four chains, each performing 3000 draws, of which the first 1000 are used as burn-in. We place a normal prior with mean 0 and standard deviation 5 on each coefficient, which ensures the prior is weakly informative. We use following diagnostics: trace plots, effective sample size, and Rhat to assess convergence and detect potential sampling issues.

### B. Results and Discussion

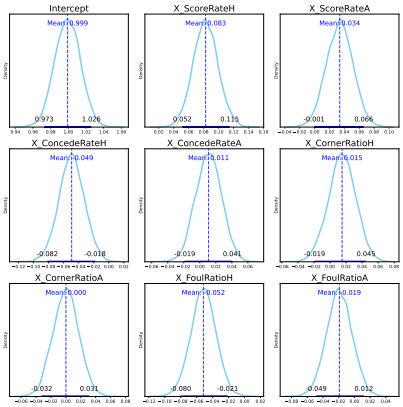


Fig. 1: Posterior estimation of coefficients, along with their means and 94% HDI.

Based on the posterior distributions, we find that the average score per game is 2.7. *X_ScoreRateH* has a positive effect on the total number of goals scored. On the other hand, *X_ConcedeRateH* and *X_FoulRatioH* show a negative effect, and higher values of these decrease the total number of goals scored.

For all the other variables, we are uncertain about their effect because their 94% HDI includes zero, we do not have enough evidence to say whether they increase or decrease goal counts.

*1) Autocorrelaion:* Analyzing autocorrelation across different chains, we observe that at lag 1, autocorrelation drops below 0.05. This indicates that the samples are independent from each other and that our sampling is efficient.

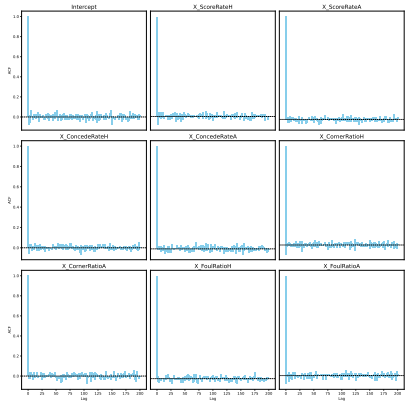


Fig. 2: Autocorrelation for one chain to lag = 200.

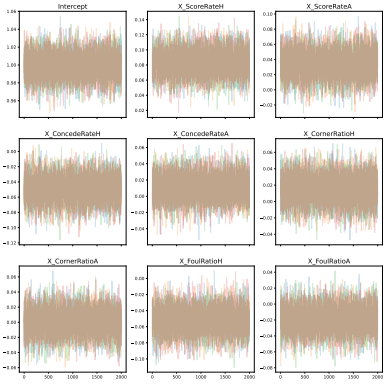


Fig. 3: Trace plot for all the parameters.

*2) Trace:* Analyzing the trace plots, we can see that all coefficients show good mixing across chains. The chains are overlapping, exploring the same regions of the parameter space, and are not stuck in any specific area. From this, we can assume that all chains have converged well.

### C. ESS and Rhat

The R-hat values for all model parameters are between 0.9998 and 1.0008. This means that all MCMC chains have converged well and are sampling from the same distribution, so the results are reliable. Also, the effective sample size (ESS) is above 7000 for each parameter, with some values even exceeding 8000. This means that a high number of effectively independent samples were used due to low or slightly negative autocorrelation between samples.

TABLE I: R-hat values for all model parameters.

| **Parameter** | **R-hat** | **ESS** |
|---|---|---|
| Intercept | 1.0008 | 8912 |
| X_ScoreRateH | 1.0001 | 7600 |
| X_ScoreRateA | 0.9999 | 7695 |
| X_ConcedeRateH | 1.0000 | 8638 |
| X_ConcedeRateA | 1.0002 | 7225 |
| X_CornerRatioH | 1.0007 | 8461 |
| X_CornerRatioA | 1.0003 | 8702 |
| X_FoulRatioH | 0.9998 | 8637 |
| X_FoulRatioA | 1.0004 | 8829 |

## III. Laplace Approximation

### A. Methodology

To fit the model using the Laplace approximation, we minimize the negative log-posterior. Its minimum gives the MAP estimate:

$$\hat{\theta} = \arg\min_{\theta} \left[ -\log p(\theta \mid y) \right]$$

The gradient of the negative log-posterior is:

$$\frac{\partial \left( -\log p(\theta \mid y) \right)}{\partial \theta} = X^{\top}(\lambda - y) + \frac{\theta}{\sigma^2}$$

The Hessian is:

$$\frac{\partial^2 \left( -\log p(\theta \mid y) \right)}{\partial \theta^2} = X^{\top} \operatorname{diag}(\lambda) X + \frac{1}{\sigma^2} I$$

Once we find the MAP estimate $\hat{\theta}$, we approximate the posterior as a Gaussian:

$$p(\theta \mid y) \approx \mathcal{N}(\hat{\theta}, H^{-1})$$

### B. Results and Discussion

From Figure 4, we observe that the means of the parameters match those from MCMC, and the 94% HDIs are also similar. The only difference is that the posterior from the Laplace approximation is a normal posterior by construction, while the MCMC estimates gives the exact posterior shape.

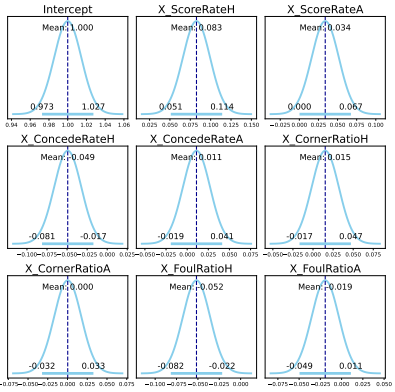


Fig. 4: Posterior estimation of coefficients using the Laplace approximation.

To make point predictions using the Laplace approximation, we first sample 1000 values from the multivariate normal distribution defined by the MAP estimate and the covariance matrix from the Laplace approximation. For each sampled parameter vector, we calculate the linear predictor by multiplying the test features with the parameter vector. Then, we exponentiate to get the Poisson rate $\lambda$. This gives us an empirical distribution of predicted goal counts for each test sample. Depending on the loss function, we choose different statistics from this distribution as point predictions. To minimize squared error, we use the mean of the predicted values. To minimize absolute error, we use the median. To maximize accuracy, we use the mode. The results on the test set are shown on Figure 5.

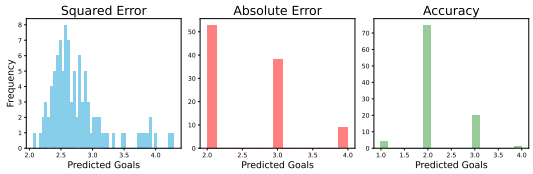


Fig. 5: Point estimation using different loss functions.

## IV. Conclusion

We confirmed convergence of MCMC chains using standard diagnostics and analyzed the posterior coefficients. The Laplace approximation gave similar results while being more computationally efficient. We also used the approximate posterior to generate point predictions under different loss functions, highlighting how predictive decisions depend on the chosen loss.