MATH 5600 Midterm Paper

Predicting Justice Scores of EU Administrative Regions Using Bayesian Logit Normal Regression

I. Coddington
Georgetown University
Washington, DC

A. Koutromanos
Georgetown University
Washington, DC

1 Introduction

Logit normal regression is commonly used to model proportion data which is bounded between 0 and 1. The response is transformed via the logit function and assumed to follow a normal distribution. A Bayesian approach to logit-normal regression extends this framework by incorporating prior distributions on model parameters, which allows for the incorporation of domain knowledge. In this paper, we apply Bayesian logit normal regression to World Justice Project (WJP) Access to Justice Survey data, which contains over 25,000 responses representing the proportion of justice needs that were met given an individual's nontrivial legal problem. Specifically, we implement a Bayesian Logit Normal Regression to predict the justice score, which is computed using the following dimensions for a given legal problem: access to information, access to representation, cost, timeliness, and fairness, and overall satisfaction with the outcome. These features are encoded as binary flags whose simple average is the justice score. Naturally, this value is between 0 and 1 inclusive, so we implement a minor transformation to shrink the range between 0 and 1. Additionally, because the justice score is computed using only 6 binary flags, its distribution contains several peaks. We address this by summarizing justice scores at the regional level. We first outline the mathematical formulation of our Bayesian model, which uses the vector or justice scores as its response and a matrix X of geographic and demographic predictors. We then describe the computational techniques used for posterior estimation, and evaluate model performance through code application.

2 Methods

2.1 Model and Likelihood

Let y_i denote the justice score of the ith region. We have responses $y_i \in (0,1), \quad i = 1, \dots, n$. We define the following logit transformation on each y_i:

$$z_i = \log\left(\frac{y_i}{1-y_i}\right)$$

where $z_i \sim \mathcal{N}(x_i^\top \beta, \sigma^2)$, x_i is a vector of features, and β is a vector of regression coefficients. We posit that the logit transformation of the justice score has a linear relationship with predictors, such that the regression equation is expressed by

$$z_i = \log\left(\frac{y_i}{1-y_i}\right) = x_i^\top \beta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

The likelihood of z is then given by

$$L(\beta, \sigma^2 \mid z) \propto (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(z - X\beta)^\top (z - X\beta)\right\} \tag{1}$$

2.2 Prior Distribution

To ensure an objective Bayesian inference framework, we employ a full Jeffrey's prior,

$$\pi(\beta, \sigma^2) \propto 1/\sigma^{p+2},$$

where p is the number of features. Jeffrey's prior is noninformative and invariant to reparameterization. The full derivation of Jeffrey's prior from the Fisher Information matrix can be found in Appendix A.

2.3 Posterior Distributions

Multiplying the likelihood and Jeffrey's prior, we find that the joint posterior for β and σ^2, up to a constant of proportionality, is

$$\begin{aligned} p(\beta, \sigma^2 \mid z) &\propto L(\beta, \sigma^2 \mid z)\, \pi(\beta, \sigma^2) \\ &\propto (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(z - X\beta)^\top (z - X\beta)\right\} \frac{1}{\sigma^{p+2}} \\ &\propto \sigma^{-(n+p+2)} \exp\left\{-\frac{1}{2\sigma^2}(z - X\beta)^\top (z - X\beta)\right\}. \end{aligned}$$

When σ^2 is held fixed, the β-dependent part of the joint posterior is

$$p(\beta \mid \sigma^2, z) \propto \exp\left\{-\frac{1}{2\sigma^2}(z - X\beta)^\top (z - X\beta)\right\}.$$

This is the kernel of a multivariate normal density. Completing the square shows that

$$\beta \mid \sigma^2, z \sim \mathcal{N}\left(\hat{\beta},\ \sigma^2 (X^\top X)^{-1}\right),$$

where $\hat{\beta} = (X^\top X)^{-1} X^\top z$.

To obtain the marginal posterior for σ^2, we integrate the joint posterior with respect to β:

$$p(\sigma^2 \mid z) \propto \int_{\mathbb{R}^p} p(\beta, \sigma^2 \mid z)\, d\beta.$$

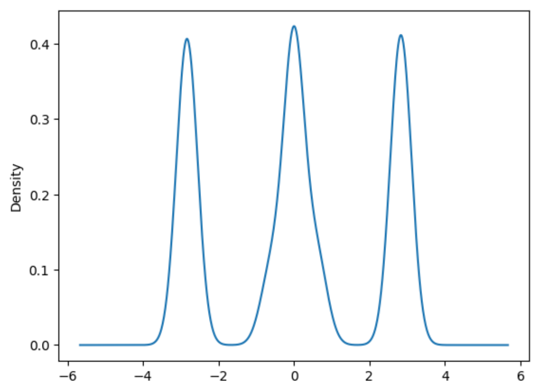
Evaluation of this integral gives the kernel of an inverse-gamma density with shape $\frac{n}{2}$ and scale $\frac{SSE}{2}$. Thus,

$$\sigma^2 \mid z \sim \text{Inverse-Gamma}\left(\frac{n}{2}, \frac{SSE}{2}\right),$$

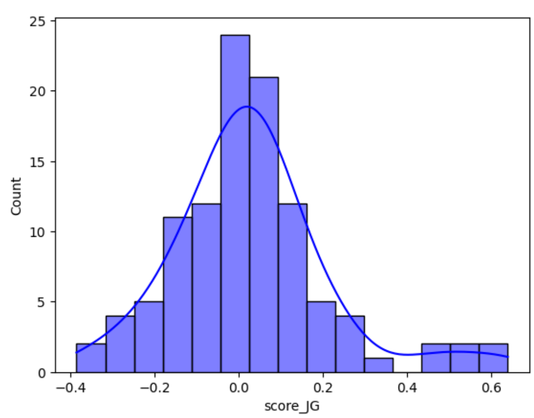
with $= z^\top \left[I - X(X^\top X)^{-1} X^\top\right] z$.

2.4 Data Cleaning and Preparation

Before applying Bayesian inference to Access to Justice data, we consider the distribution of our vector of y_i's and impose necessary transformations on our data to address the assumptions of logit normal regression. The transformed response vector of z_i's is distributed as follows:

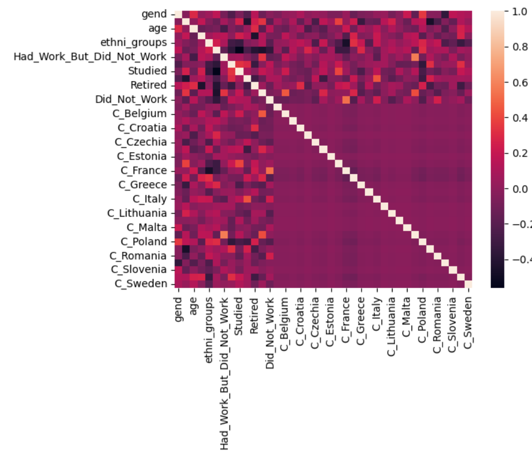


Since this distribution is trimodal with peaks at -3, 0, and 3, it should not be used directly for Bayesian Logit Normal Regression. We instead group the data by administrative region, of which there are 107 in WJP's EU Access to Justice Module. The distribution of z_i's then has the following shape:



Next we consider the features of interest. We use age and level of education, which are is treated as a continuous values, employment status, gender, country, administrative region, urbanity, and ethnic minority status. For each of these features, a value of 98 or 99 represents a response of "Don't Know" and "No Answer", respectively, and thus is ommitted from our sample of interest. Categorical variables with only two levels are treated as binary variables, and categorical variables with more than $k = 2$ levels are transformed to $k - 1$ indicator variables. This is the case for employment status and country.

It is also useful to consider potential interaction between variables. We consider correlation among our features:



While there are some indicators that appear moderately correlated, no relationship is significant enough to introduce interaction features into our model.

We now move forward in applying our model using Python.

3 Simulation Study and Application

3.1 Procedure

To obtain posterior samples from our analytical derivations of the posterior distributions of β and σ^2, we first sample σ^2 from its inverse-gamma posterior. We then sample β from its conditional normal posterior given σ^2. Posterior predictive means are then computed, where μ is given by

$$\mu = X_{\text{test}}\beta.$$

Finally we generate justice score predictions from the logit normal distribution, where

$$y_{\text{pred}} = \frac{1}{1 + \exp(-\mu_{\text{pred}})}.$$

These steps are implemented using Numpy and SciPy for efficient matrix operations.

3.2 Model Performance and Evaluation

To assess model performance, we compute evaluation metrics on the test set, which is a random draw of twenty percent of the full sample which was not used to construct the model. The chosen metrics reflect the goal of assessing the overall prediction accuracy of the model. We choose the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-Squared Score, and Posterior Predictive p-value. The R-Squared Score measures the proportion of variance in justice score that is explained by the model and is calculated as follows:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}i)^2}{\sum i(y_i - \bar{y})^2}.$$

R-squared provides an intiuitive measure of model fit, since a value of 1 indicated perfect prediction, a value of 0 means

Table 1: Model performance metrics for Bayesian Logit-Normal Regression.

Metric	Value
Mean Squared Error (MSE)	0.1226
Root Mean Squared Error (RMSE)	0.3502
Mean Absolute Error (MAE)	0.2408
R^2 Score	-2.6506
Posterior Predictive p-value (p_B)	0.4660

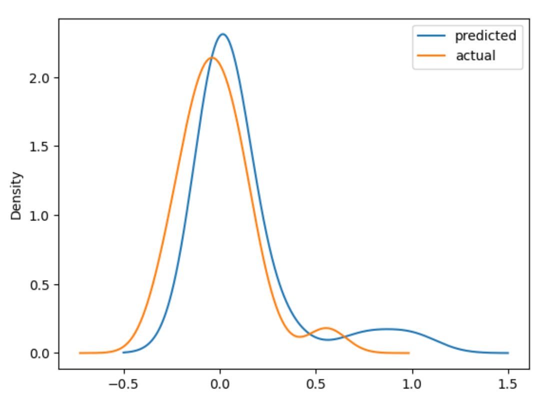
the model performs exactly the same as predicting the mean, and a negative value indicates that the model is worse than predicting the mean. The negative R-squared score suggests that our model struggles to explain variance in the justice score. This may be due to high variability in individual experiences that country- and demographic-level features alone cannot capture.

Posterior predictive checks are conducted by generating simulated replicated data and comparing its test statistics to the observed data. The posterior predicted p-value is computed as:

$$p_B = \mathbb{P}[T(y^{\text{rep}}, \theta) \geq T(y, \theta) \mid y],$$

where $T(y, \theta)$ is the sum of squared differences. Using 500 Monte Carlo samples, we estimate p_B via Monte Carlo integration. Our posterior predictive p value is close to 0.5, which suggests the the model generates plausible replicated data, indicating a reasonable fit.

To visually inspect model fit, we compare the predicted justice scores with actual test set scores using a kernel density estimate (KDE) plot:



The density plot shows that the predicted distribution broadly follows the actual distribution. However, it sppears over-smoothed, with fewer extreme values than the actual density. This suggests that variance in justice scored is underestimated, potentially due to the assumption that the logit transformation of justice score comes from a normal distribution. It could also be the case that geographic and demographic features are simply not strongly predictive of the accessibility of justice in a given region.

4 Discussion

4.1 Model Performance

Our results semonstrate that Bayesian Logit-Normal regression provides a principles probabilistic framework for predicting regional justice score based on demographic and geographic factors. However, the negative R squared score suggests that the model struggles to explain variability in justice score. This indicates that justice accessibility may not be strongly determined by the chosen features, or that a linear relationship in logit space does not fully capture the complexity of regional justice needs.

Despite the low R-squared, the posterior predictive (p_B = 0.466) suggests that the model generates plausible replicated data. Since a posterior predictive p-value close to 0.5 indicates that the observed data are not systematically different from simulated predictions, this result suggests that the Bayesian Logit-Normal model is not severely misspecified. However, it may not fully capture the underlying structure of the data.

4.2 Transformation Effects and Normality Assumption

The justice score is transformed using the logit function, assuming that the transformed values follow a normal distribution. However, our exploratory analysis shows that even after transformation, the justice scores exhibit a slightly multimodal distribution in grouped data. This suggests that normality assumptions may not fully hold, potentially reducing model accuracy.

4.3 Feature Selection

The model relies on demographic and geographic factors (e.g., age, education, country, urban status), but legal system characteristics, socioeconomic inequalities, and policy structures may play a larger role in justice accessibility. Since such predictors are not included in the current model, important sources of variability in justice score remain unexplained.

4.4 Comparison of Regional vs. Individual-Level Data

Our approach involved summarizing justice scores at the regional level to ensure that the logit-transformed scores followed a distribution closer to normality. This helped satisfy normality assumption at the cost of potentially masking within region disparities in access to justice.

4.5 Future Directions and Alternative Approaches

Future iterations of the model should explore legal system characteristics (e.g., procedural fairness, court efficiency, legal representation rates) as predictors. Additionally, Instead of assuming a normal distribution in logit space, we could explore models that assume a Beta distribution for justice scores, avoiding logit transformation entirely.

A Model Derivations

A.1 Jeffrey's Prior

Jeffrey's prior is given by

$$\pi(\theta) \propto \sqrt{|\mathcal{I}(\theta)|}$$

where $\mathcal{I}(\theta)$ is the Fisher information matrix for the parameter vector $\theta = (\beta, \sigma^2)$. We assume that after the logit transformation, the response follows a normal distribution:

$$z_i = \log\left(\frac{y_i}{1-y_i}\right) \sim \mathcal{N}(x_i^\top \beta, \sigma^2).$$

For a dataset with n observations, the likelihood is then

$$L(\beta, \sigma^2 \mid z) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(z - X\beta)^\top (z - X\beta)\right).$$

Taking the log of the likelihood function:

$$\log L(\beta, \sigma^2 \mid z) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(z - X\beta)^\top (z - X\beta).$$

The Fisher information matrix is

$$\mathcal{I}(\beta, \sigma^2) = -\mathbb{E}\left[\frac{\partial^2 \log L(\beta, \sigma^2)}{\partial(\beta, \sigma^2)^2}\right].$$

The first derivative of the log-likelihood with respect to β is

$$\frac{\partial \log L}{\partial \beta} = \frac{1}{\sigma^2} X^\top (z - X\beta).$$

Taking the second derivative:

$$\frac{\partial^2 \log L}{\partial \beta \partial \beta^\top} = -\frac{1}{\sigma^2} X^\top X.$$

The Fisher information block for β is therefore

$$\mathcal{I}_{\beta\beta} = \frac{1}{\sigma^2} X^\top X.$$

The first derivative of the log-likelihood with respect to σ^2 is

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(z - X\beta)^\top (z - X\beta).$$

Taking the second derivative:

$$\frac{\partial^2 \log L}{\partial \sigma^2 \partial \sigma^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}(z - X\beta)^\top (z - X\beta).$$

Taking the expectation over the distribution of z, we use:

$$\mathbb{E}[(z - X\beta)^\top (z - X\beta)] = n\sigma^2.$$

The Fisher information component for σ^2 is therefore

$$\mathcal{I}_{\sigma^2\sigma^2} = \frac{n}{2\sigma^4}.$$

Since β and σ^2 are independent in the likelihood, their mixed derivatives are zero. The full information matrix is:

$$\mathcal{I}(\beta, \sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2} X^\top X & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}.$$

The determinant of $\mathcal{I}(\beta, \sigma^2)$ is

$$\left|\mathcal{I}(\beta, \sigma^2)\right| = \left|\frac{1}{\sigma^2} X^\top X\right| \times \left|\frac{n}{2\sigma^4}\right|.$$

Since the determinant of $X^\top X$ is $|X^\top X|$, we get

$$\left|\mathcal{I}(\beta, \sigma^2)\right| = \frac{|X^\top X|}{\sigma^{2p}} \times \frac{n}{2\sigma^4}. = \frac{n|X^\top X|}{2\sigma^{2p+4}}. \tag{2}$$

Taking the square root to obtain Jeffrey's prior:

$$\pi(\beta, \sigma^2) \propto \sqrt{\frac{n|X^\top X|}{2\sigma^{2p+4}}}.$$

Since $|X^\top X|$ is constant with respect to σ^2, we simplify:

$$\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^{p+2}}.$$

A.2 Joint Posterior for β and σ^2

Using Bayes' rule, the joint posterior is:

$$p(\beta, \sigma^2 \mid z) \propto L(\beta, \sigma^2 \mid z)\pi(\beta, \sigma^2). \tag{3}$$

Substituting the likelihood and Jeffrey's prior:

$$p(\beta, \sigma^2 \mid z) \propto (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(z - X\beta)^\top (z - X\beta)\right) \frac{1}{\sigma^{p+2}}.$$

Simplifying:

$$p(\beta, \sigma^2 \mid z) \propto \sigma^{-(n+p+2)} \exp\left(-\frac{1}{2\sigma^2}(z - X\beta)^\top (z - X\beta)\right).$$

A.3 Conditional Posterior of β Given σ^2

The part of the joint posterior that depends on β is:

$$\exp\left(-\frac{1}{2\sigma^2}(z - X\beta)^\top (z - X\beta)\right). \tag{4}$$

Expanding the quadratic term:

$$(z - X\beta)^\top (z - X\beta) = z^\top z - 2\beta^\top X^\top z + \beta^\top X^\top X\beta.$$

Rearrange:

$$(z - X\beta)^\top (z - X\beta) = \beta^\top X^\top X\beta - 2\beta^\top X^\top z + z^\top z.$$

Thus, the β dependent part of the posterior is:

$$\exp\left(-\frac{1}{2\sigma^2}\left(\beta^\top X^\top X\beta - 2\beta^\top X^\top z\right)\right). \tag{5}$$

We complete the square in β.

$$\beta^\top X^\top X\beta - 2\beta^\top X^\top z.$$

Define:

$$\hat{\beta} = (X^\top X)^{-1} X^\top z.$$

Rewriting,

$$\beta^\top X^\top X\beta - 2\beta^\top X^\top z = (\beta - \hat{\beta})^\top X^\top X(\beta - \hat{\beta}) - \hat{\beta}^\top X^\top X\hat{\beta}. \tag{6}$$

Thus, the exponential term becomes:

$$\exp\left(-\frac{1}{2\sigma^2}(\beta - \hat{\beta})^\top X^\top X(\beta - \hat{\beta})\right).$$

This is the kernel of a multivariate normal density with mean $\hat{\beta}$ and covariance matrix $\sigma^2(X^\top X)^{-1}$. Since the terms that do not depend on β are absorbed into the normalization constant, we obtain:

$$\beta \mid \sigma^2, z \sim \mathcal{N}\left((X^\top X)^{-1} X^\top z, \sigma^2 (X^\top X)^{-1}\right).$$

A.4 Marginal Posterior of σ^2

To obtain the marginal posterior of $\sigma^2 \mid z$, β is integrated out of the joint posterior:

$$p(\sigma^2 \mid z) = \int_{\mathbb{R}^p} p(\beta, \sigma^2 \mid z)\, d\beta.$$

From the previous result,

$$p(\beta, \sigma^2 \mid z) \propto \sigma^{-(n+p+2)} \exp\left(-\frac{1}{2\sigma^2}(z - X\beta)^\top (z - X\beta)\right).$$

Rewrite $(z - X\beta)^\top (z - X\beta)$ by completing the square:

$$(z - X\beta)^\top (z - X\beta) = (z - X\hat{\beta})^\top (z - X\hat{\beta}) + (\beta - \hat{\beta})^\top X^\top X(\beta - \hat{\beta}).$$

Thus, we can separate the quadratic form:

$$\exp\left(-\frac{1}{2\sigma^2}(z - X\beta)^\top (z - X\beta)\right) = \exp\left(-\frac{1}{2\sigma^2}(z - X\hat{\beta})^\top (z - X\hat{\beta})\right) \times \exp\left(-\frac{1}{2\sigma^2}(\beta - \hat{\beta})^\top X^\top X(\beta - \hat{\beta})\right).$$

Integrating out β yields

$$p(\sigma^2 \mid z) \propto \sigma^{-(n+2)/2} \exp\left(-\frac{(z - X\hat{\beta})^\top (z - X\hat{\beta})}{2\sigma^2}\right).$$

This matches the kernel of an inverse gamma distribution:

$$p(\sigma^2 \mid z) \propto (\sigma^2)^{-(n/2+1)} \exp\left(-\frac{SSE}{2\sigma^2}\right),$$

where $SSE = (z - X\hat{\beta})^\top (z - X\hat{\beta})$. Therefore,

$$\sigma^2 \mid z \sim \text{Inverse-Gamma}\left(\frac{n}{2}, \frac{SSE}{2}\right).$$

B Replication Code

Our replication code can be found in the file alongside this manuscript.