SMOKING PREDICTION USING HIERARCHICAL MODEL

— Project Report —
Applied Bayesian Data Analysis

Mohammad Ibtesum Sakib, Rayen Werda

March 14, 2025

TU Dortmund University

1 Introduction

Smoking remains a significant global public health problem, and more insight into the demographic and socioeconomic determinants that influence smoking status is required in order to design effective, targeted interventions. The issue of the prediction of smoking behavior is addressed in this study by means of individual-level factors—age, gender, and income—controlling for regional variation in smoking prevalence. Our research question is, **"Can we predict whether an individual is likely to smoke based on income and other factors?"** To tackle this, three complementary modeling approaches are proposed: a Bayesian hierarchical model with random region-level (**Midlands and East Anglia, Scotland, South East, South West, The North, Wales**) intercepts, and a Bayesian hierarchical model that uses random slope and another with varying slope and intercept. These models aim to disentangle the complex interplay of individual and contextual determinants of smoking behavior.

2 Data

The dataset originates from kaggle, a platform primarely used for data science competitions, where participants can compete with each other to create the best models for solving various challenges. The Dataset used in this study "Smoking Dataset from UK" originates from a Survey on smoking habits from the United Kingdom. The data set can be used for analyzing the demographic characteristics of smokers and types of tobacco consumed. The dataset comprises 1691 observations on 12 variables that are a mix of 9 categorical and 3 numerical variables.

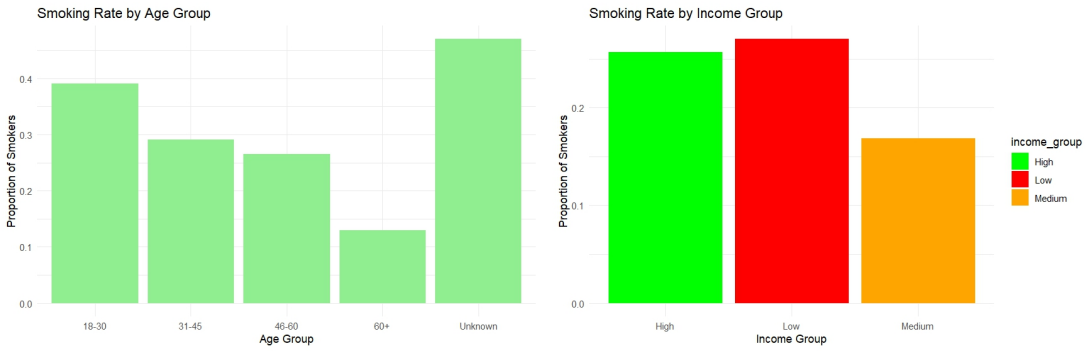

Figure 1: Demographic distribution of smokers: (Left) Age group distribution, (Right) Income group distribution.

The first graph charts the rate of smoking in different age groups, where the highest rate of smokers can be seen among the 18-30 group, and decreasing rates of smoking in the 31-45 and 46-60 groups. The smallest rate of smoking is observed for the 60+ group, perhaps because they are more afflicted by health ailments or due to differences in habits between generations. Interestingly, the "Unknown" category has the highest smoking percentage, which might be a sign of missing or unclassified information.

The second graph indicates the smoking percentage among income groups and finds that the lowest-income group has the highest smoking percentage, followed by the high-income group, while the medium-income group has the lowest smoking percentage. This pattern is reflective of the fact that smoking is prevalent among lower-income groups, which might be due to stress, affordability, or social reasons. However, the relatively high prevalence of smoking among affluent individuals indicates that there are some other variables besides economic status, such as lifestyle or preference, in operation.

Overall, the two graphs suggest that there is a correlation between age, income, and smoking patterns, with young and poor individuals more likely to smoke. That there is a high rate of smoking in the "Unknown" category in the age split suggests potential data quality problems that will have to be removed in data preprocessing part.

2.1 Data preprocessing

In the initial preprocessing phase, unnecessary columns such as index values, smoking type, and the number of cigarettes smoked were removed because of the irrelevance for our analysis. The cases containing "refused" or "unknown" responses in categorical variables such as nationality, ethnicity, and gross income were then excluded to maintain data integrity. Then, all categorical variables including gender, marital status, highest qualification, nationality, ethnicity, gross income, and region were converted to factor variables.

3 Models

In this report, we analyze the factors that influence smoking using Bayesian hierarchical regression models. The motivation behind this study is to understand how various socioeconomic features, such as age, region, income, and education, contribute to the probability of smoking. In this report, three distinct Bayesian hierarchical regression models were generated, each incorporating smoking as a dependent variable.

3.1 Model 1 Bayesian Hierarchical Model

Model 1 applies a Bayesian hierarchical model to predict smoking status (binary: smoker vs. non-smoker) as a function of individual-level predictors (age, gender, and gross income) with regional(**Midlands and East Anglia, Scotland, South East, South West, The North, Wales**) variation accounted for through a random intercept term. The hierarchical model allows baseline smoking probabilities to vary by region, acknowledging geographic differences in smoking prevalence. Fixed effects on income, gender, and age were assigned weakly informative priors

$$\text{Normal}(0, 0.5)$$

, with a more exact prior

$$\text{Normal}(0, 0.02)$$

for the age coefficient to reflect a conservative estimate of its magnitude. Student-t priors

$$\text{Student-t}(3, 0, 5)$$

and

$$\text{Student-t}(3, 0, 2.5)$$

were assigned for the intercept and regional random effects to be outlier robust. 4 chains of 2,000 iterations (1,000 warmup) with Bernoulli likelihood were utilized to fit the model so probabilistic inference is possible for how demographic and regional influences combined produce smoking behavior. This is a balance between interpretability and flexibility that distinguishes regional variation from global trends estimated across the population. The formula is:

$$\begin{aligned}\text{smoke} \sim\ & \text{age} + \text{gender} + \text{marital\_status} + \text{highest\_qualification} \\ & + \text{nationality} + \text{ethnicity} + \text{gross\_income} \\ & + (1 \mid \text{region})\end{aligned}$$

3.2 Model 2

Model 2 further extends the hierarchical framework by introducing region-specific variation in the effects of highest qualification on smoking behavior. Unlike Models 1 which model regional differences through a random intercept, Model 2 replaces the random intercept with a region-specific random slope for highest qualification. This means that the effect of educational attainment on smoking varies across regions, allowing for more nuanced regional differences in how education influences smoking status.the formula is:

$$\begin{aligned}\text{smoke} \sim\ & \text{age} + \text{gender} + \text{marital\_status} + \text{highest\_qualification} \\ & + \text{nationality} + \text{ethnicity} + \text{gross\_income} \\ & + (0 + \text{highest\_qualification} \mid \text{region})\end{aligned}$$

the model Regression Coefficients are shown in table 2.

3.3 Model 3

Model 3 builds upon Model 1 by incorporating additional individual-level predictors: marital status, highest qualification, nationality, and ethnicity. The model is also a Bayesian hierarchical model, predicting smoking status (binary: smoker vs. non-smoker) while accounting for regional variation through a random intercept and random slope term for highest qualification across the six regions. The formula is:

$$\begin{aligned}\text{smoke} \sim\ & \text{age} + \text{gender} + \text{marital\_status} + \text{highest\_qualification} \\ & + \text{nationality} + \text{ethnicity} + \text{gross\_income} \\ & + (1 + \text{highest\_qualification} \mid \text{region})\end{aligned}$$

the model Regression Coefficients are shown in table 3.

4 Priors

Model 1 Priors are **fixed effects** on income, gender, and age were assigned weakly informative priors

$$\text{Normal}(0, 0.5)$$

, with a more exact prior

$$\text{Normal}(0, 02)$$

for the age coefficient to reflect a conservative estimate of its magnitude. Student-t priors

$$\text{Student-t}(3, 0, 5)$$

and

$$\text{Student-t}(3, 0, 2.5)$$

were assigned for the intercept and regional random effects to be outlier robust. Similar priors were assigned for **Model 2 and 3** in addition to an

$$\text{LKJ}(2)$$

prior that is placed on the correlation between the random intercept and slope, ensuring a reasonable structure for the regional effects for **Model 3**.

5 Code

The statistical analysis for this project was conducted within the R programming language environment, version 4.4.3 . R served as the platform for data processing, statistical modeling, and visualization tasks, owing to its extensive range of packages and libraries tailored for various analytical needs. Key libraries employed in this project include brms for fitting Bayesian multilevel models using Stan, loo for efficient leave-one-out cross-validation grid, gridextra and ggplot for visualisation,rstan for interfacing with Stan,a probabilistic programming language for Bayesian inference, and tidyr for data manipulation and transformation tasks. These libraries collectively provided the necessary tools and functionalities for conducting rigorous statistical analysis, model fitting, visualization, and data manipulation, thereby contributing to the successful execution of the project objective.

5.1 Models explanation

We present Bayesian hierarchical logistic regression models aimed at predicting smoking behavior based on individual-level demographic and socioeconomic factors. The models consider independent variables such as age, gender, gross income, marital status, highest qualification, nationality, ethnicity, and region. The dependent variable in all models is smoking status (binary: smoker vs. non-smoker). The first model specifies a random intercept for region, allowing baseline probabilities of smoking to vary geographically. The

second model introduces region-specific random slopes for the highest qualification, which allow for regional variation in the effect of education on smoking behavior. The third model specifies both random intercepts and slopes for region, providing a more flexible specification for capturing regional heterogeneity. They employ Bayesian approaches to model the probability of smoking, incorporating uncertainty and prior knowledge on the relations of the predictors to smoking.

Model 1: Bayesian Hierarchical Logistic Regression Model with Random Intercept for Region

```
model1 ← brm(
  formula = smoke ∼ age + gender + gross_income + marital_status +
            highest_qualification + nationality + ethnicity + (1 |
               region),
  data = data,
  family = bernoulli(),
  prior = c(
    prior(normal(0, 0.5), class = "b"),
    prior(normal(0, 0.02), class = "b", coef = "age"),
    prior(student_t(3, 0, 5), class = "Intercept"),
    prior(student_t(3, 0, 2.5), class = "sd")
  ),
  chains = 4,
  iter = 2000,
  warmup = 1000,
  cores = 4,
  seed = 123
)
```

Model 2: Bayesian Hierarchical Logistic Regression Model with Random Slopes for Highest Qualification by Region

```
model2 ← brm(
  formula = smoke ∼ age + gender + gross_income + marital_status +
            nationality + ethnicity + highest_qualification +
            (0 + highest_qualification | region),
  data = data,
  family = bernoulli(),
  prior = c(
    prior(normal(0, 0.5), class = "b"),
    prior(normal(0, 0.02), class = "b", coef = "age"),
    prior(student_t(3, 0, 5), class = "Intercept"),
    prior(student_t(3, 0, 2.5), class = "sd", group = "region")
  ),
  chains = 4,
  iter = 2000,
  warmup = 1000,
  cores = 4,
  seed = 123
)
```

Bayesian Hierarchical Logistic Regression Model with Random Intercept and Slope for Highest Qualification by Region

```
model3 ← brm(
  formula = smoke ~ age + gender + marital_status + highest_
     qualification +
            nationality + ethnicity + gross_income +
            (1 + highest_qualification | region),
  data = data,
  family = bernoulli(),
  prior = c(
    prior(normal(0, 0.5), class = "b"),
    prior(normal(0, 0.02), class = "b", coef = "age"),
    prior(student_t(3, 0, 5), class = "Intercept"),
    prior(student_t(3, 0, 2.5), class = "sd", group = "region"),
    prior(lkj(2), class = "cor")
  ),
  chains = 4,
  iter = 2000,
  warmup = 1000,
  cores = 4,
  seed = 123
)
```

6 Convergence diagnostics

We conducted convergence diagnostics using the Bayesian framework implemented in the brms package for all the models. The convergence diagnostics were performed on the Bayesian hierarchical logistic regression models designed for predicting smoking behavior. The model specifications comprised a Bernoulli likelihood function and weakly informative priors for all fixed effects, with normal priors (mean = 0, standard deviation = 0.5) for the coefficients and student-t priors (df = 3, mean = 0, scale = 5) for the intercept. For the random effects, we used student-t priors (df = 3, mean = 0, scale = 2.5) for the standard deviations and an LKJ prior (shape = 2) for the correlation matrix in the model with varying intercepts and slopes. The models were fitted to the data with 2000 iterations per chain and a warm-up period of 1000 iterations using 4 CPU cores. The potential scale reduction factor (R^) approached 1 for all parameters, indicating convergence. Overall, the convergence diagnostics validated that the model achieved reliable parameter estimation and offered valid inference for predicting smoking based on the specified predictors.

6.1 Posterior Predictive Checks

In the context of our analysis, posterior predictive checks were used on our Bayesian hierarchical logistic regression models to evaluate the models' ability to replicate the observed data. The posterior predictive checks were performed using the `pp_check()` function in R, which compares the observed smoking status data to simulated results generated from

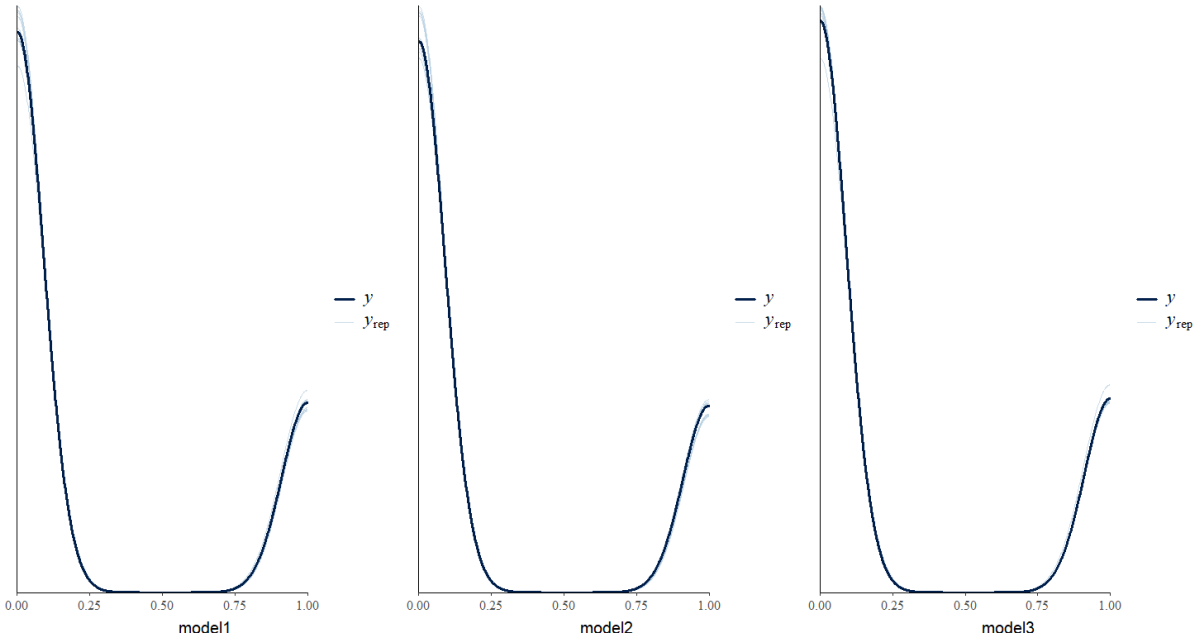

Figure 2: posterior predictive checks

the posterior predictive distribution. For each model (`model1`, `model2`, and `model3`), we generated posterior predictive distributions and visualized them alongside the observed data. The resulting plots allow us to evaluate how well each model captures the variability and patterns in the data.

7 Model Comparaison

In this section, we compare the performance of three different bayesian hierarchical logistic regression models developed for predicting smoke. The models under consideration are denoted as **model1**,**model2** and **model3**.

7.1 Leave One Out Cross Validation LOO

LOO provides an estimate of predictive performance by calculating the out-of-sample prediction error for each data point, and then averaging these errors. This is essentially testing the model on unseen data, giving an idea of its ability to generalize. The result of loo are as follows: **model1:**

Table 3: LOO results for Model1

Metric	Estimate (SE)
elpd_loo	-818.4 (20.6)
p_loo	25.3 (0.9)
looic	1637.6 (41.1)

All Pareto k-values are good (k < 0.7), indicating reliable LOO estimates.
model2:

Table 4: LOO results for Model2

Metric	Estimate (SE)
elpd_loo	-816.8 (20.9)
p_loo	41.9 (1.7)
looic	1633.5 (41.7)

All Pareto k-values are good (k < 0.7), indicating reliable LOO estimates.
model3:

Table 5: LOO results for Model3

Metric	Estimate (SE)
elpd_loo	-816.4 (20.8)
p_loo	41.5 (1.7)
looic	1632.8 (41.7)

All Pareto k-values are good (k < 0.7), indicating reliable LOO estimates.

LOO comparaison

Table 6: LOO Comparison Results

Model	elpd_diff (se_diff)
Model 3	0.0 (0.0)
Model 2	-0.4 (0.8)
Model 1	-2.0 (4.0)

the results of the **LOO** comparaison shows that **model3** achieves the best results.

8 Limitations and potential improvements

The models employed in regression can fail to capture the nature of complexity present in the relationship between smoking and socioeconomic factors. Investigating more ad-

vanced modeling techniques, e.g., deep learning networks, provides a route to potentially enhancing prediction performance. In addition, the dataset contains "unknown" or "refused" responses in categorical variables such as income, ethnicity, and nationality that were excluded during preprocessing. This could lead to bias in the analysis since some subpopulations might be underrepresented, and therefore the model performance and interpretability would be compromised.

9 Conclusion

The aim of this study was to predict smoking behavior using three Bayesian hierarchical logistic regression models with individual-level factors (age, gender, income, education, etc.) and regional variation. After model fitting, Model 3, with both random intercepts and slopes for education by region, performed the best. It had the lowest LOO Information Criterion (LOOIC) and the highest effective number of parameters `p_loo`, which represented the best predictive accuracy. Model 2, with random slopes for education by region, performed slightly less well, and Model 1, with only a random intercept, had the worst accuracy. Lastly, Model 3 is the most accurate, yielding a more precise and flexible prediction that accounts for regional variation in baseline smoking prevalence as well as the impact of education.

10 Appendix

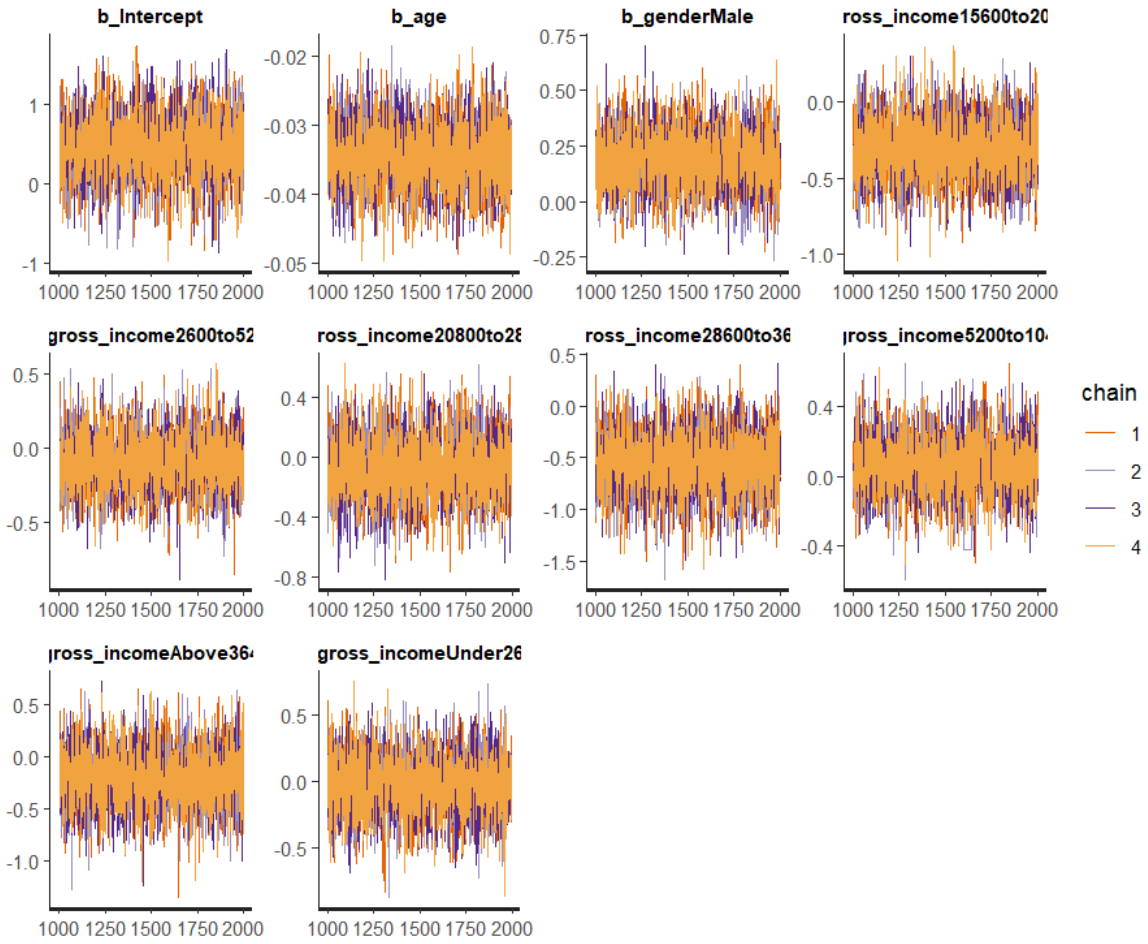

Figure 3: traceplot model1

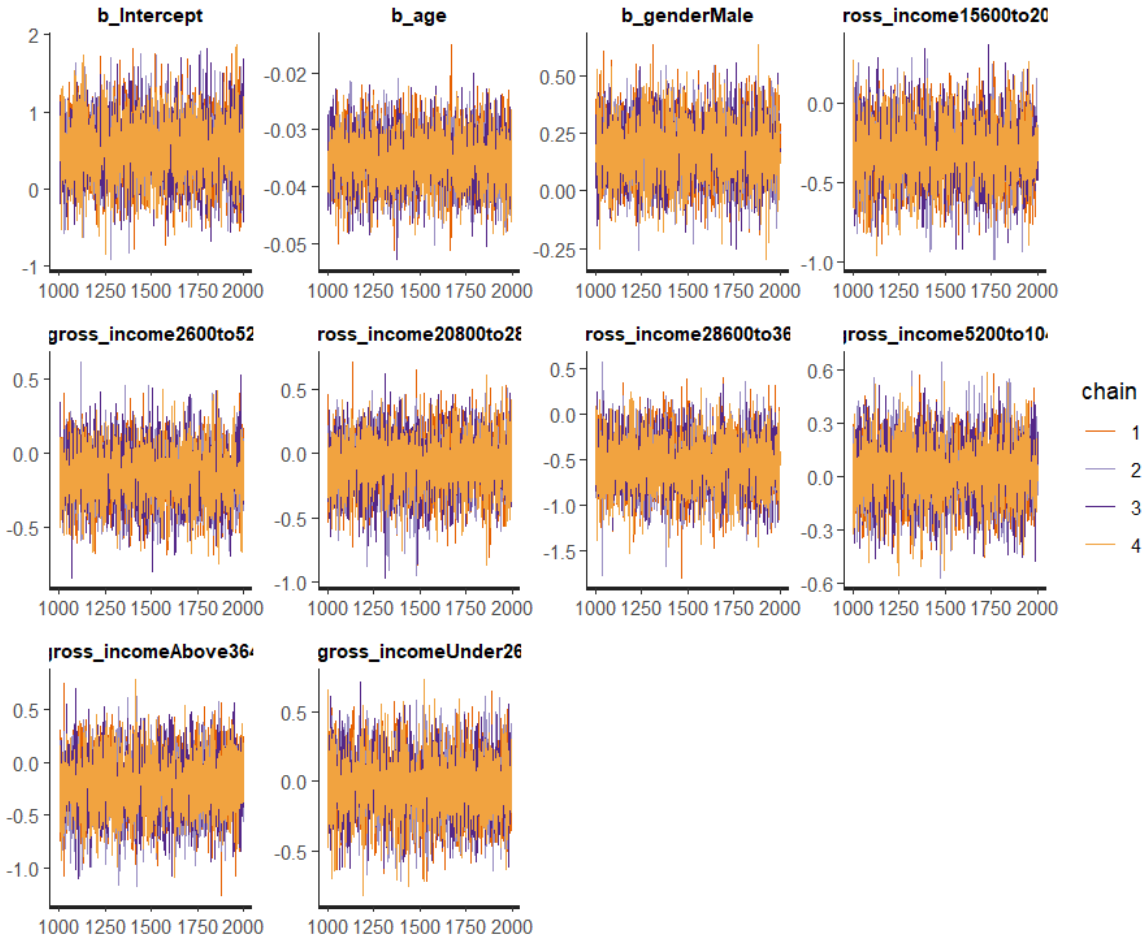

Figure 4: traceplot model2

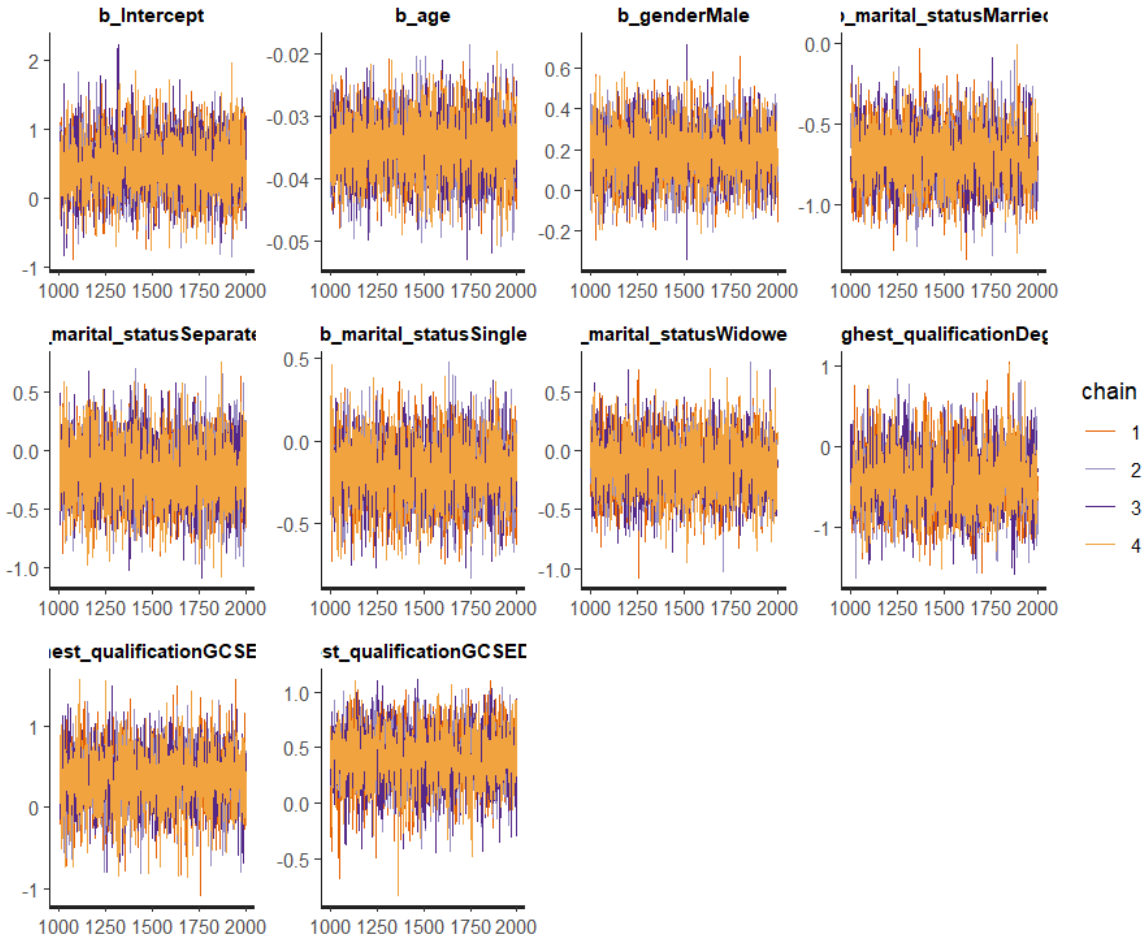

Figure 5: traceplot model3

Table 1: Regression Coefficients for the Bayesian Logistic Model with varying slope

Variable	**Estimate (95% CI)**
Intercept	0.52 (-0.29, 1.33)
age	-0.04 (-0.04, -0.03)
genderMale	0.18 (-0.08, 0.44)
gross_income15600to20800	-0.31 (-0.69, 0.07)
gross_income2600to5200	-0.15 (-0.54, 0.21)
gross_income20800to28600	-0.09 (-0.52, 0.33)
gross_income28600to36400	-0.54 (-1.15, 0.04)
gross_income5200to10400	0.03 (-0.31, 0.36)
gross_incomeAbove36400	-0.22 (-0.76, 0.34)
gross_incomeUnder2600	-0.04 (-0.48, 0.41)
marital_statusMarried	-0.70 (-1.05, -0.35)
marital_statusSeparated	-0.18 (-0.73, 0.36)
marital_statusSingle	-0.19 (-0.57, 0.20)
marital_statusWidowed	-0.15 (-0.59, 0.31)
nationalityEnglish	-0.02 (-0.29, 0.24)
nationalityIrish	0.69 (0.02, 1.38)
nationalityOther	-0.46 (-1.02, 0.09)
nationalityScottish	0.29 (-0.18, 0.75)
nationalityWelsh	-0.13 (-0.69, 0.44)
ethnicityBlack	0.04 (-0.68, 0.74)
ethnicityChinese	-0.15 (-0.93, 0.59)
ethnicityMixed	0.22 (-0.59, 1.03)
ethnicityWhite	0.27 (-0.24, 0.82)
highest_qualificationDegree	-0.45 (-1.19, 0.37)
highest_qualificationGCSEDCSE	0.30 (-0.43, 0.96)
highest_qualificationGCSEDOLevel	0.35 (-0.18, 0.86)
highest_qualificationHigherDSubDegree	-0.02 (-0.61, 0.54)
highest_qualificationNoQualification	0.42 (-0.06, 0.85)
highest_qualificationONCDBTEC	0.01 (-0.65, 0.65)
highest_qualificationOtherDSubDegree	0.29 (-0.30, 0.83)

Table 2: Regression Coefficients for the Bayesian Logistic Model with varying slope and intercept

Variable	**Estimate (95% CI)**
Intercept	0.48 (-0.32, 1.29)
age	-0.04 (-0.04, -0.03)
genderMale	0.18 (-0.08, 0.44)
marital_statusMarried	-0.70 (-1.04, -0.35)
marital_statusSeparated	-0.17 (-0.70, 0.35)
marital_statusSingle	-0.19 (-0.56, 0.19)
marital_statusWidowed	-0.14 (-0.60, 0.32)
highest_qualificationDegree	-0.42 (-1.13, 0.36)
highest_qualificationGCSEDCSE	0.35 (-0.33, 0.99)
highest_qualificationGCSEDOLevel	0.39 (-0.11, 0.86)
highest_qualificationHigherDSubDegree	0.03 (-0.51, 0.57)
highest_qualificationNoQualification	0.46 (-0.01, 0.88)
highest_qualificationONCDBTEC	0.05 (-0.58, 0.68)
highest_qualificationOtherDSubDegree	0.33 (-0.23, 0.85)
nationalityEnglish	-0.01 (-0.28, 0.25)
nationalityIrish	0.68 (-0.00, 1.35)
nationalityOther	-0.45 (-1.01, 0.11)
nationalityScottish	0.19 (-0.39, 0.68)
nationalityWelsh	-0.12 (-0.73, 0.51)
ethnicityBlack	0.03 (-0.68, 0.77)
ethnicityChinese	-0.16 (-0.96, 0.61)
ethnicityMixed	0.23 (-0.55, 1.04)
ethnicityWhite	0.27 (-0.24, 0.81)
gross_income15600to20800	-0.31 (-0.72, 0.08)
gross_income2600to5200	-0.15 (-0.52, 0.22)
gross_income20800to28600	-0.09 (-0.52, 0.33)
gross_income28600to36400	-0.54 (-1.11, 0.01)
gross_income5200to10400	0.03 (-0.31, 0.36)
gross_incomeAbove36400	-0.22 (-0.77, 0.31)
gross_incomeUnder2600	-0.04 (-0.50, 0.41)

Table 7: Regression Coefficients for the Bayesian Logistic Model with Varying Intercept

Variable	**Estimate (95% CI)**
Intercept	0.45 (-0.35, 1.21)
age	-0.03 (-0.04, -0.02)
genderMale	0.19 (-0.07, 0.45)
gross_income15600to20800	-0.30 (-0.69, 0.08)
gross_income2600to5200	-0.09 (-0.46, 0.28)
gross_income20800to28600	-0.07 (-0.49, 0.35)
gross_income28600to36400	-0.56 (-1.16, 0.01)
gross_income5200to10400	0.06 (-0.28, 0.39)
gross_incomeAbove36400	-0.20 (-0.75, 0.33)
gross_incomeUnder2600	-0.03 (-0.47, 0.40)
marital_statusMarried	-0.68 (-1.02, -0.32)
marital_statusSeparated	-0.16 (-0.69, 0.42)
marital_statusSingle	-0.17 (-0.54, 0.21)
marital_statusWidowed	-0.14 (-0.60, 0.33)
highest_qualificationDegree	-0.46 (-0.89, -0.02)
highest_qualificationGCSEDCSE	0.43 (-0.05, 0.88)
highest_qualificationGCSEDOLevel	0.41 (0.05, 0.77)
highest_qualificationHigherDSubDegree	0.01 (-0.46, 0.50)
highest_qualificationNoQualification	0.47 (0.08, 0.85)
highest_qualificationONCDBTEC	0.07 (-0.45, 0.63)
highest_qualificationOtherDSubDegree	0.34 (-0.16, 0.82)
nationalityEnglish	-0.02 (-0.28, 0.25)
nationalityIrish	0.72 (0.07, 1.39)
nationalityOther	-0.44 (-1.00, 0.13)
nationalityScottish	0.20 (-0.41, 0.69)
nationalityWelsh	-0.11 (-0.72, 0.47)
ethnicityBlack	0.06 (-0.66, 0.76)
ethnicityChinese	-0.11 (-0.92, 0.68)
ethnicityMixed	0.27 (-0.54, 1.02)
ethnicityWhite	0.24 (-0.25, 0.76)

References

Auguie, B. (2017). Gridextra: Miscellaneous functions for "grid" graphics [R package version 2.3]. https://CRAN.R-project.org/package=gridExtra

Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Wickham, H. (2019). Tidyverse: Easily install and load the 'tidyverse' [R package version 1.3.0]. https://CRAN.R-project.org/package=tidyverse

Vehtari, A., Gelman, A., Gabry, J., & Yao, Y. (2020). Loo: Efficient leave-one-out cross-validation and waic for bayesian models [R package version 2.4.1]. https://CRAN.R-project.org/package=loo

Wickham, H. (2021). Tidyr: Tidy messy data [R package version 1.2.0]. https://CRAN.R-project.org/package=tidyr

Team, S. D. (2022). Rstan: R interface to stan [R package version 2.21.5]. https://mc-stan.org/

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://doi.org/https://doi.org/10.1007/978-3-662-67526-7