A longitudinal analysis study on women's contraception

longitudinal final project

Kindle Zhang

December 2024

1 Abstract

In this paper, I will use the Contraception dataset from the R package to study the factors influencing women's use of contraceptive methods. These data, collected from the 1988 Bangladesh Fertility Survey[1], provide information on contraceptive use among women in urban and rural areas. I will employ both the GLMM model and the marginal model to investigate this question and compare their performances. Finally, whether in terms of goodness of fit or predictive ability, the GLMM outperforms the marginal model in this question. Briefly, holding other factors constant, older women are more likely to use contraception, younger women are less likely, women with more children are more likely to use contraception and women living urban are more likely to use contraception.

2 Key words

GLMM, marginal model, GEE 1.0, GEE 1.5 , validation, likelilood ratio test

3 Introduction

The following are the variables I will use in the analysis. The outcome variable is use, while the other variables are covariates:

- **use**: Contraceptive use at the time of the survey (binary categorical).

- **district**: Identifier for each district.

- **livch**: Number of living children at the time of the survey (ordered factor: 0, 1, 2, 3+).

- **age**: Age of the woman at the time of the survey, centered around the mean (numeric).

- **urban**: Region of residence (factor: urban or rural).

I will use the GEE method to obtain the marginal model and the ML method to obtain the GLMM model. I will compare their corresponding parameter estimates and inference.

I will split the dataset into a test group and a training group in a 2:8 ratio. After building the models, I will use cross-validation to compare the predictive performance of the two models. And choose the model with better performance as my final model to explain women's using on contraception.

4 Methods

4.1 EDA

Before we do the EDA, we try to clean the data. In the output, the variable **use** is a binary variable, while the covariates **livch** and **urban** are categorical variables. The variable **age** is a continuous variable related to time. To facilitate better understanding, I adjusted the starting point of the **age** variable to 0 and applied rounding.

4.1.1 use vs livch or urban

If we try to find a relationship between **use** and **livch** or **use** and **urban**, we can use a chi-square test in each **district** group. Here are the results of p-values for chi-square tests and I pit them in a histogram graph: **Figure**1 and **Figure**2

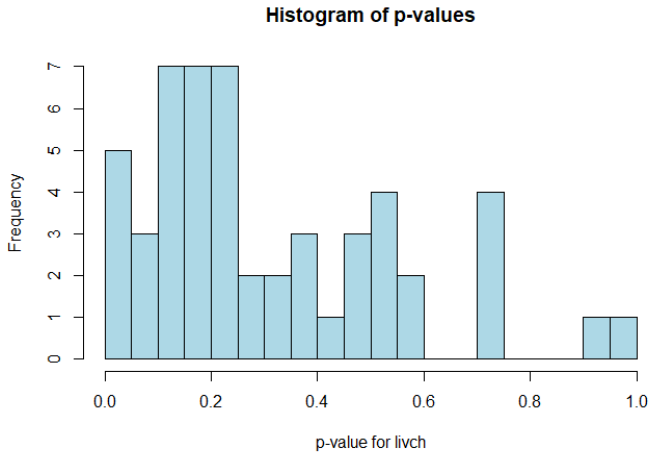

Figure 1: P-value for use vs livch histogram

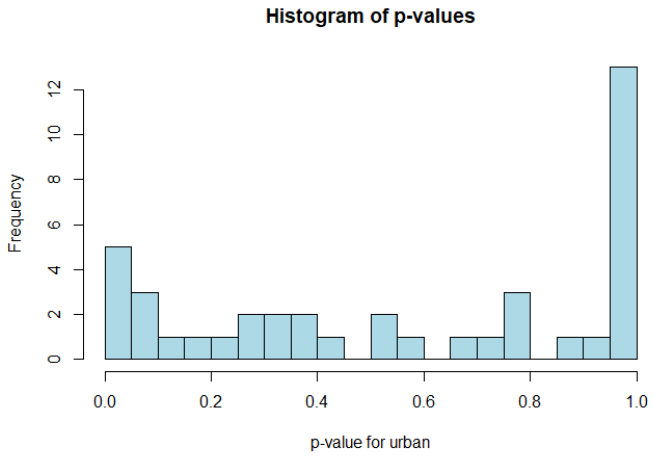

Figure 2: P-value for use vs urban histogram

We can find that the p values of most districts are higher than 0.05, which means that **use** and **covariate** are independent in most districts.

However, if we treat all the data as independent and conduct a chi-square test, we find that **use** and the two covariates are highly correlated.(p-value is 6e-12 and 1e-10 namely) I suspect this might be because the sample size within each district is too small, making the chi-square test results unreliable in each group. Nevertheless, we can still include these two variables in our model.

4.1.2 use vs age

In the meanwhile, I plot a longitudinal response data for a sample of 10 districts from the K=61 in this study. Here is the plot: **Figure**3 The red point means using the contraception method.

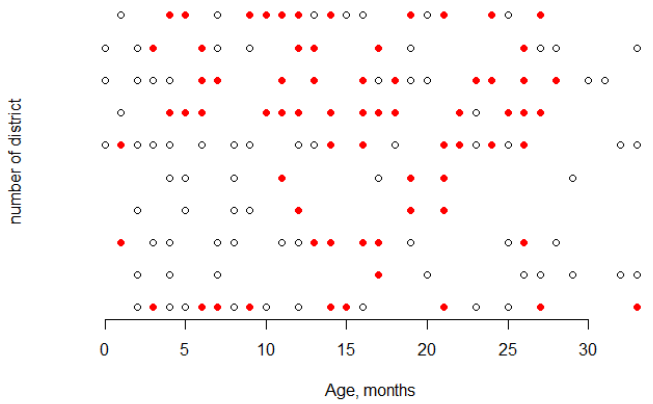

Figure 3: the contraception use in 10 districts

From the images, it is difficult to directly perceive the relationship between the variables **age** and **use**. Therefore, we should include age in the model for further investigation.

4.2 Marginal Model: GEE

Recall the corvariate I will put in my model: **livch**, **urban** and **age**, consider the model:

$$\text{logit}\,\Pr(Y_{ki}|\mathbf{X}_{ki}) = \beta_0 + \beta_1 \text{Age}_{ki} + \beta_2 \text{livch1}_{ki} + \beta_3 \text{livch2}_{ki} + \beta_4 \text{livch3}_{ki} + \beta_5 \text{Urban}_{ki}$$

In this equation, $\exp(\beta_2)$ represents the odds ratio for livch1 compared to the reference category of livch (e.g., livch0). This shows how the odds of the outcome change when livch1 is present, relative to the baseline livch0, holding all other variables constant.

4.2.1 GEE 1.0

We attempted to use three types of working covariance matrices for the analysis: independent, exchangeable, and auto-regressive with MV=1. However, the **VC_GEE_covlag** argument has exceeded the **MAX_COVLAG** limit for the number of rows in your working covariance matrix, so we only get first two results.

To be clearer to what is going on, we can use a geeglm() function instead of gee() function so that we can get all three models. Actually, the geeglm function use a GEE 1.5 methods, but I still want to use it here to compare three models and two methods (GEE 1.0 and GEE 1.5)

working independence Model Results					
independent	Estimate	Model Error	Robust Error	Estimate_glm	robust_SE_glm
(Intercept)	−1.354	0.131	0.150	-1.3542	0.15003
age	−0.014	0.008	0.008	-0.0138	0.00831
urban	0.808	0.117	0.199	0.8076	0.19856
livch1	1.042	0.168	0.191	1.0417	0.19089
livch2	1.240	0.187	0.170	1.2401	0.16962
livch3+	1.081	0.191	0.210	1.0815	0.21024

Figure 4: independent matrix model

working exchangeable Model Results					
exchangeable	Estimate	Model Error	Robust Error	Estimate_glm	robust_SE_glm
Intercept	−1.392	0.155	0.151	-1.3917	0.15109
Age	−0.015	0.008	0.008	-0.0152	0.00823
livch1	0.738	0.127	0.176	0.7373	0.17607
livch2	1.060	0.167	0.200	1.0604	0.19962
livch3+	1.271	0.186	0.163	1.2713	0.16332
Urban	1.163	0.191	0.209	1.1638	0.20906

Figure 5: exchangeable matrix model

	Estimate	Std.err	Wald	Pr(>|W|)
(Intercept)	-1.333	0.151	78.22	0.000
age	-0.013	0.008	2.65	0.103
urban	0.780	0.186	17.50	0.000
livch1	1.049	0.200	27.55	0.000
livch2	1.162	0.171	46.29	0.000
livch3+	1.045	0.213	24.18	0.000

Figure 6: AR1 matrix model

The noticeable difference between the model-based and robust standard errors for the estimates under the working exchangeable and independence correlation structure suggests that these two may not be the best approximation to the true dependence structure, and a different working correlation matrix might provide a better fit.

However, the overall fit of the model is strong, as the Wald p-value for each parameter is less than 0.05, regardless of the working correlation matrix used.

4.2.2 GEE 1.5

The central idea of GEE 1.5 is to replace the moment-based estimators of α with a different estimator based on an additional set of estimating equations.

In this part of the analysis, I introduce a new variable, **density**, which is a cluster-specific binary variable indicating population density. A value of 1 represents densely populated areas, while 0 represents sparsely populated areas. My goal is to investigate the effect of population **density** on whether women use contraception, after adjusting for other variables. To account for the correlation within clusters, I will use two types of working dependence matrices: (1) an exchangeable structure with a common α for both crowded and uncrowded districts, and (2) an exchangeable structure with separate α for crowded and uncrowded districts, allowing for different correlations between observations in crowded and uncrowded districts. This approach enables a more nuanced understanding of the impact of population density within different correlation structures.

Whether using GEE 1.5 with a common α or different α, the p-value for the coefficient of **density** is greater than 0.05, indicating it is not significant. Furthermore, as shown in Figure 9, there is no evidence that the correlation parameter differs between crowded and uncrowded districts. Therefore,

GEE Model Results for Contraception Use

Variable	Estimate	Standard Error	Wald Statistic	P-value	Confidence Interval (Lower)	Confidence Interval (Upper)
(Intercept)	−1.387	0.177	61.656	0.000	−1.733	−1.041
age	−0.015	0.008	3.391	0.066	−0.031	0.001
urban	0.737	0.176	17.532	0.000	0.392	1.082
livch1	1.060	0.200	28.168	0.000	0.669	1.452
livch2	1.271	0.163	60.714	0.000	0.951	1.591
livch3+	1.164	0.209	31.145	0.000	0.755	1.572
density	−0.009	0.174	0.003	0.959	−0.350	0.332

Figure 7: GEE 1.5 with common α

```
 Coefficients:
             estimate  san.se     wald        p
(Intercept) -1.38950 0.17680 61.76916 3.89e-15
age         -0.01519 0.00822  3.41176 6.47e-02
urban        0.74027 0.17566 17.75965 2.51e-05
livch1       1.05964 0.20020 28.01617 1.20e-07
livch2       1.26968 0.16336 60.40861 7.66e-15
livch3+      1.16498 0.20894 31.08819 2.47e-08
density     -0.00636 0.17395  0.00134 9.71e-01
```

Figure 8: GEE 1.5 with separate α

we don't have to retain **density** in our model.

4.3 GLMM Model: ML or REML

Unlike the marginal model, the mixed model explicitly includes random effects in the model, allowing for a clearer exploration of the specific dependence relationships between clusters.

Consider the (more general) **logistic-Normal GLMM**:

$$\text{logit}\,\mu_{ki} = \mathbf{X}_{ki}\boldsymbol{\beta}^* + \mathbf{Z}_{ki}\boldsymbol{\gamma}_k$$

where the $\boldsymbol{\gamma}_k$ are i.i.d. $\text{MVN}(0, G(\alpha))$.

4.3.1 GLMM with random intercept

In this project, if we make $Z_{ki} = 1$, then we will get a GLMM model with random intercept. The exact model is:

```
 Estimated Correlation Parameters:
        estimate san.se   wald        p
alpha:1   0.0696 0.0193 13.016 0.000309
alpha:2  -0.0134 0.0322  0.174 0.676371
```

Figure 9: separate α

$$\text{logit}\Pr(Y_{ki}|\mathbf{X}_{ki}, u_k) = \beta_0 + \gamma_{0k} + \beta_1 \text{Age}_{ki} + \beta_2 \text{livch1}_{ki} + \beta_3 \text{livch2}_{ki} + \beta_4 \text{livch3}_{ki} + \beta_5 \text{Urban}_{ki}$$

where $\gamma_{0k} \sim Normal(0, \sigma_\gamma^2)$

Before we make a regression, we should determine the best number of nodes, the result shows that using a GH method is welcomed although the returns quickly diminish as the number of nodes increases. I decide nAGQ = 5 finally. This can increase the accuracy of the approximation to the integrated likelihood.

Comparison of Estimates Across nAGQ Values

Parameter	nAGQ=1	nAGQ=5	nAGQ=10	nAGQ=25
β*_0	−1.4684	−1.4691	−1.4691	−1.4691
β*_1	−0.0159	−0.0159	−0.0159	−0.0159
β*_2	0.7853	0.7848	0.7848	0.7848
β*_3	1.1085	1.1089	1.1089	1.1089
β*_4	1.3373	1.3377	1.3377	1.3377
β*_5	1.2181	1.2189	1.2189	1.2189
alpha	0.4943	0.4989	0.4989	0.4989

Figure 10: glmm beta with different nAGQ

And the result of the model:

```
Random effects:
 Groups   Name        Variance Std.Dev.
 district (Intercept) 0.249    0.499
Number of obs: 1548, groups:  district, 60

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.46908    0.15443   -9.51  < 2e-16 ***
age         -0.01592    0.00893   -1.78    0.075 .
urban        0.78478    0.13388    5.86  4.6e-09 ***
livch1       1.10894    0.17614    6.30  3.1e-10 ***
livch2       1.33765    0.19654    6.81  1.0e-11 ***
livch3+      1.21887    0.20220    6.03  1.7e-09 ***
```

Figure 11: glmm with random intercept result

4.3.2 GLMM with random intercept/slope

Similarly, if we make Z_{ki} = age, which is a time dependent variable, then we will get a GLMM model with random intercept and slope. The exact model is:

$$\text{logit}\,\Pr(Y_{ki}|\mathbf{X}_{ki}, u_k) = \beta_0 + \gamma_{0k} + \beta_1 \text{Age}_{ki} + \beta_2 \text{livch1}_{ki} + \beta_3 \text{livch2}_{ki} + \beta_4 \text{livch3}_{ki} + \beta_5 \text{Urban}_{ki} + \beta_6 \text{Age}\gamma_{1k}$$

where $\gamma_k \sim MVN(0, G(\alpha))$

So we can get the result of a GLMM with random slope and intercept:

```
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 district (Intercept) 0.24775  0.4977
          age         0.00027  0.0164   -0.24
Number of obs: 1548, groups:  district, 60

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.46641    0.16038   -9.14  < 2e-16 ***
age         -0.01629    0.00945   -1.72    0.085 .
urban        0.78886    0.13428    5.87  4.2e-09 ***
livch1       1.11025    0.17824    6.23  4.7e-10 ***
livch2       1.33679    0.19738    6.77  1.3e-11 ***
livch3+      1.22115    0.20299    6.02  1.8e-09 ***
```

Figure 12: glmm with random intercept result

After that, I want to explore whether a random slope is necessary in this situation, so I try to compare the LRT statistic to a 1:1 mixture of $chi^2(1)$ and $chi^2(2)$ distribution. The result shows that the test statistics is 0.639 and the p-value is 0.575 which is larger than the 0.05, meaning we don't have to put a random slope variable in the model. In the meanwhile, we can also get the same conclusion by checking the AIC of two models, the random intercept model AIC is 1931 which is smaller that intercept/slope model's AIC, 1934.

4.4 validation

When we want to perform validation, there are two key aspects to discuss: how well the model fits the data and how good the model is at making predictions. In this part, the marginal model we choose GEE 1.5 with exchangeable dependence matrix and GLMM random intercept model.

4.4.1 goodness of fit

Since the GEE method does not provide a log-likelihood, we cannot use AIC to compare the goodness-of-fit between the GLMM and GEE models. Instead, we use residual analysis plots as a basis for comparison.

Residual plots allow us to visually assess the model fit and detect potential issues such as systematic

patterns, heteroscedasticity, or clustering effects that might indicate differences in how the two models capture the data structure. This approach provides an alternative way to evaluate model performance without relying on likelihood-based metrics.

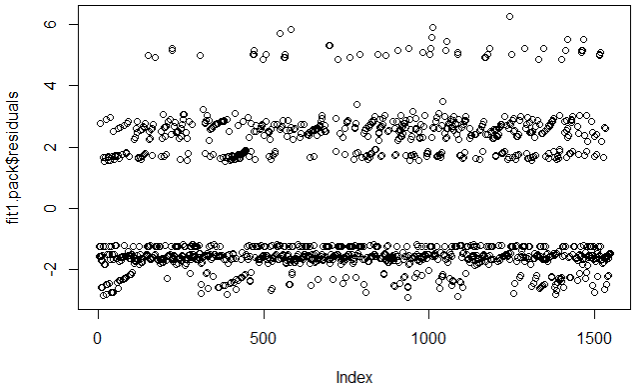

Figure 13: Marginal model residuals plot

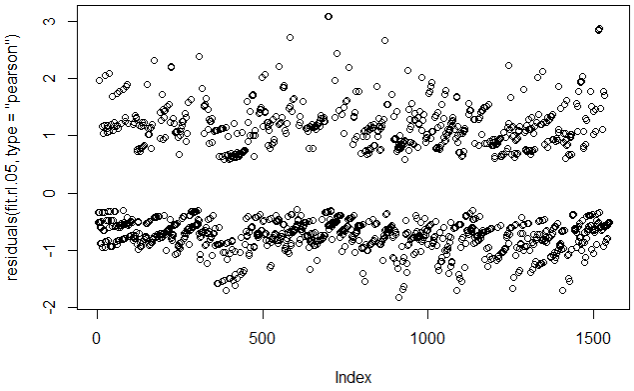

Figure 14: GLMM model residuals plot

Based on the two residual plots, it can be observed that the residuals for GlMM models are mostly distributed between -2 and 2. The range of residuals in Marginal model is wider than GLMM's, which is around -4 to 4. What's more, the marginal model shows a small number of outliers, whereas the GLMM model has almost no outliers. From this perspective, the GLMM demonstrates better performance compared to the marginal model.

4.4.2 prediction

Next, I want to compare the predictive abilities of the two models by using the test data that was set aside at the beginning of the project. The resulting ROC curves are shown below:

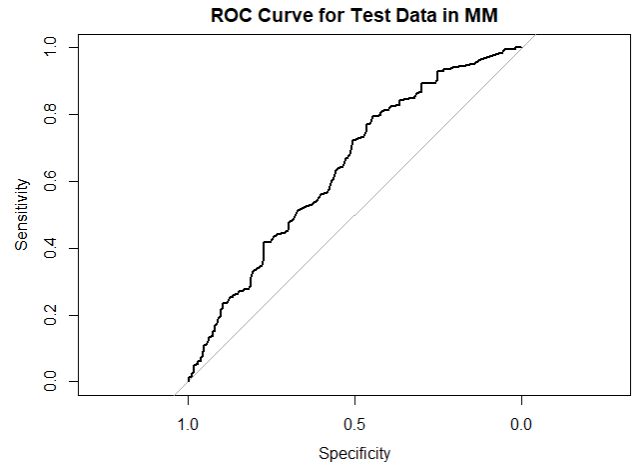

Figure 15: Marginal model ROC curve

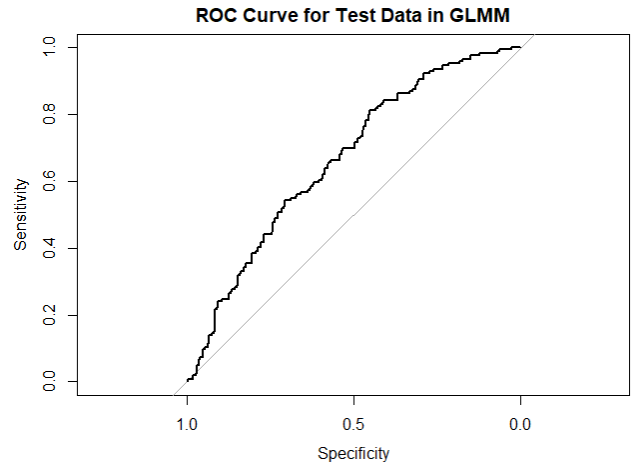

Figure 16: GLMM model ROC curve

According to the plot and the result, we know the AUC for marginal model is 0.6386 which is smaller

than AUC for marginal model, 0.6621. Therefore, the GLMM model has a better prediction ability. However, the AUC of both model are less than 0.7 which means a poor performance in prediction generally.

In summary, whether in terms of goodness of fit or predictive ability, the GLMM outperforms the marginal model.

4.4.3 compare coefficients with GLMM and MM

Let's compare the parameters of the two models under the exchangeable correlation matrix. We can observe that the significance of each variable is largely consistent across the models, the coefficients show only minor differences, and the signs are consistent.

	Estimate	Std.err	Pr(>|W|)	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-1.392	0.151	0.000	-1.469	0.154	0.000
age	-0.015	0.008	0.066	-0.016	0.009	0.075
urban	0.737	0.176	0.000	0.785	0.134	0.000
livch1	1.060	0.200	0.000	1.109	0.176	0.000
livch2	1.271	0.163	0.000	1.338	0.197	0.000
livch3+	1.164	0.209	0.000	1.219	0.202	0.000

Figure 17: two model coefficients

5 Results

In conclusion, we use the GLMM random intercept model to describe the relationship between contraceptive use among women in different districts and the variables: the number of children, age, and urban residence. Below are the regression results and their interpretation: According to the **Figure**18,

$$\text{logit}\,\Pr(Y_{ki}|\mathbf{X}_{ki}, u_k) = \beta_0 + \gamma_{0k} + \beta_1 \text{Age}_{ki} + \beta_2 \text{livch1}_{ki} + \beta_3 \text{livch2}_{ki} + \beta_4 \text{livch3}_{ki} + \beta_5 \text{Urban}_{ki}$$

Random effect:

1. **Variance (0.249)**: The variance of the random intercepts across districts is 0.249. This indicates variability in the baseline log-odds of contraceptive use between districts.

Fixed effect:

```
Random effects:
 Groups   Name        Variance Std.Dev.
 district (Intercept) 0.249    0.499
Number of obs: 1548, groups:  district, 60

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.46908    0.15443   -9.51  < 2e-16 ***
age         -0.01592    0.00893   -1.78    0.075 .
urban        0.78478    0.13388    5.86  4.6e-09 ***
livch1       1.10894    0.17614    6.30  3.1e-10 ***
livch2       1.33765    0.19654    6.81  1.0e-11 ***
livch3+      1.21887    0.20220    6.03  1.7e-09 ***
```

Figure 18: glmm with random intercept result

1. **intercept(-1.46908)**: the baseline log-odds of contraceptive use for women in rural areas (urban = 0), with no children(livch = 0) and at reference age (age = 0).

2. **Age(-0.01592)**: For each one-unit increase in age, the log-odds of contraceptive use decrease by 0.01592, holding other variables constant.

3. **Urban(0.78478)**: The corresponding odds ratio is exp(0.78478) =2.19, meaning women in urban areas are approximately 2.19 times more likely to use contraception than those in rural areas(p¡0.001), holding other variables constant.

4. **Livch1(1.10894)**: Women with one child have log-odds of contraceptive use that are 1.10894 higher than women with no children, holding other variables constant.

5. **Livch2(1.33765)**: Women with one child have log-odds of contraceptive use that are 1.33765 higher than women with no children, holding other variables constant.

6. **Livch3+(1.21887)**: Women with one child have log-odds of contraceptive use that are 1.21887 higher than women with no children, holding other variables constant.

Briefly, holding other factors constant, older women are more likely to use contraception, younger women are less likely, women with more children are more likely to use contraception and women living urban are more likely to use contraception.

6 Discussion

GLMM and marginal models serve different purposes and are suited for different scenarios. GLMM is preferred when cluster-specific interpretations are needed, such as understanding district-specific contraceptive use trends, or when random effects are essential to account for unmeasured cluster-level variability. It is also the better choice when dealing with small sample sizes or unequal cluster sizes, as

it explicitly models the data structure and within-cluster correlations. Marginal models, on the other hand, are better suited for estimating population-average effects and are robust to misspecifications of the correlation structure, particularly in large samples.

In this study, GLMM likely performs better because it incorporates district-level variability through random intercepts, allowing for a better fit and more precise estimates. The likelihood-based inference in GLMM provides flexibility for model comparison, and its ability to handle cluster-specific variability and outliers further enhances its performance. Given the hierarchical structure of the data (women nested within districts), GLMM effectively captures both population-level and district-specific effects, making it a more suitable model for this analysis.

References

[1] Steele, F., Diamond, I., and Amin, S. (1996). *Immunization uptake in rural Bangladesh: a multilevel analysis*. Journal of the Royal Statistical Society, Series A, **159**: 289–299.