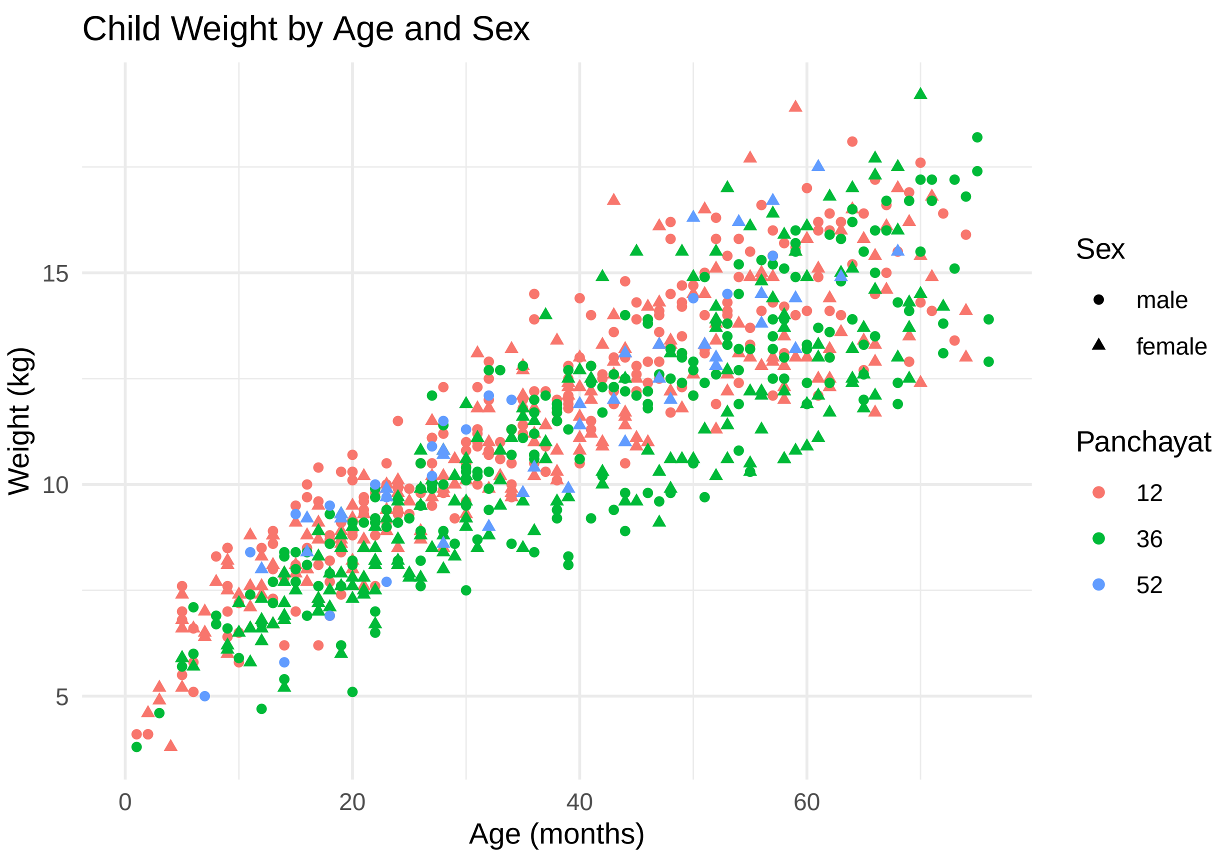
Homework Assignment 4

Greg Forkutza
Student ID: 400277514

10 December, 2023

1



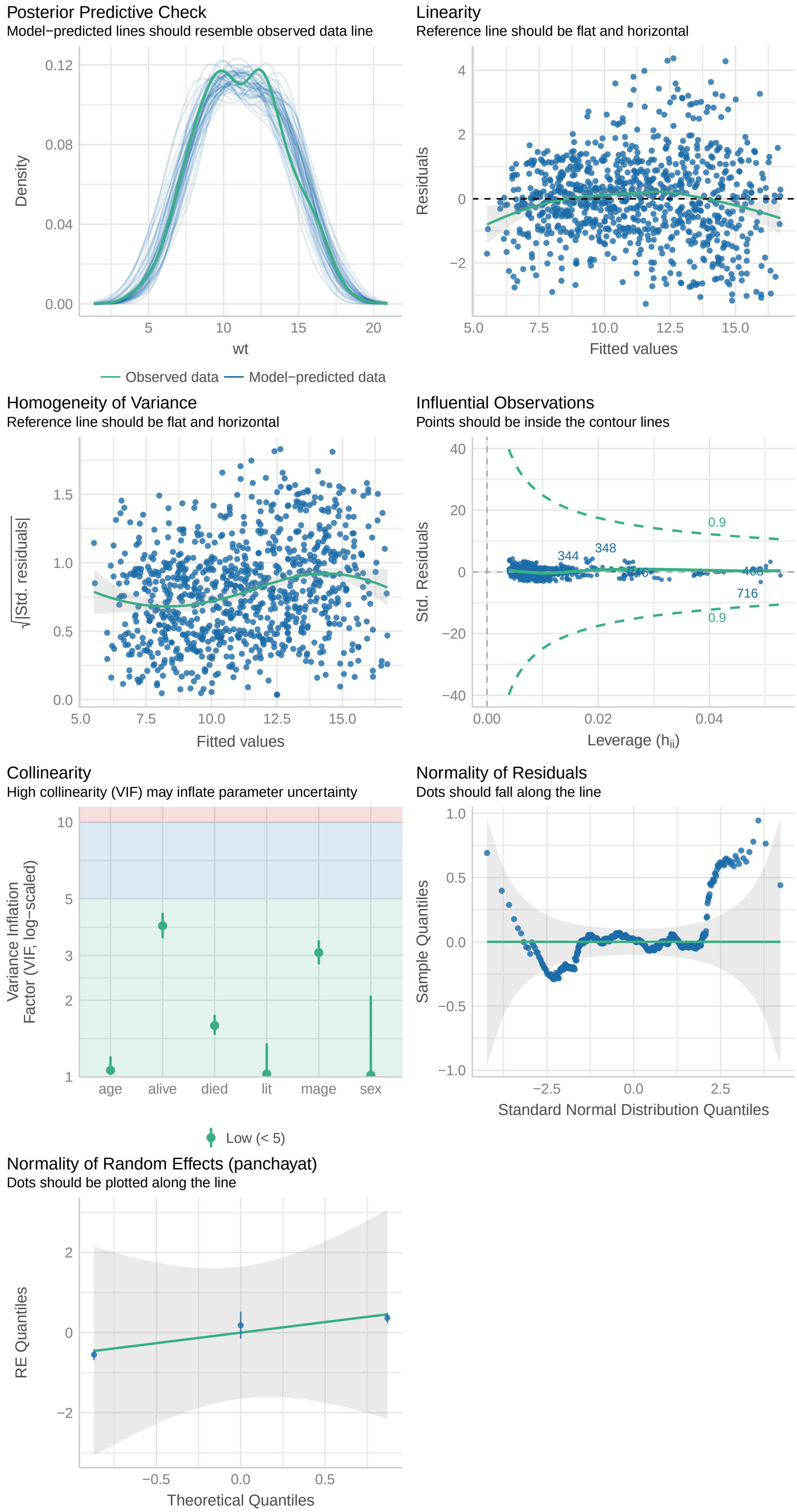
We started with a standard linear regression model using various predictors. This model showed a good fit (R-squared = 0.7843), but we found that the variables `died` (number of children who died) and `alive` (number of living children) weren't statistically significant. We then adjusted our approach by considering the `panchayat` variable, which represents different regional groups with possibly unique socioeconomic and cultural characteristics. This variable, extracted from the `id` field and treated as a factor with three levels, was included in our new model. This improved our model's fit (R-squared increased to 0.8053). Interestingly, the significance of died and alive improved, with p-values dropping to 0.0101 and 0.00012, respectively. Next, we moved to a mixed effects model using `lme4::lmer`, treating `panchayat` as a random intercept. Our goal was to capture variations across the three panchayats. We experimented with random slopes for different predictors, but only the model with `mage` (mother's age) as a random slope showed improvement (2.1% decrease in REML score). This model suggested variability in mother's age impact across regions. Using `performance::check_model`, we evaluated the mixed effects models. The diagnostic plots for both models were similar, showing a significant deviation in the tails of the residuals. The QQ plot indicated issues with the normality of residuals, consistent with the earlier linear regression models. Initially, we leaned towards the mixed effects model with the random effects term `(panchayat|mage)`. However, after further analysis using `ggeffects::ggpredict` and considering the observed data, we concluded that the model without the random slope for mother's age was more realistic. This decision was supported by an LRT test favoring this simpler model. Additionally, removing `died` from the model further improved its fit.

The final model suggests significant variability in baseline child weight across different panchayats. The residual variance of 1.710 indicates that the model doesn't fully explain the variation in child weight. Literate mothers are associated with an estimated increase of 1.038kg in child weight compared to non-literate mothers. Each additional living child is linked with a decrease in weight by approximately 0.113kg. Each additional year in the mother's age correlates with an estimated weight increase of 0.071 kg in children. This analysis indicates that mother's literacy and the number of living children have a notable impact on child weight, alongside variations linked to different panchayats.

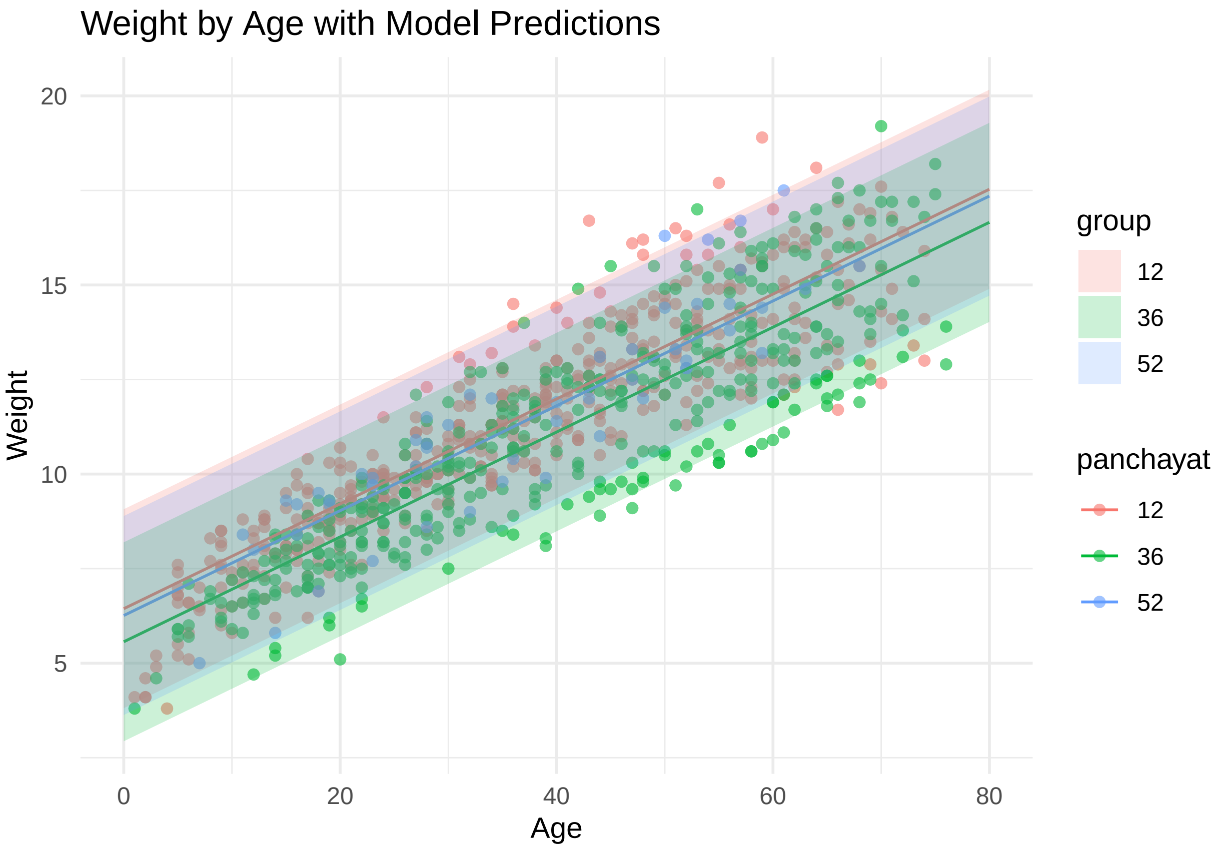
```
## Linear mixed model fit by REML ['lmerMod']
## Formula: wt ~ age + sex + lit + alive + mage + died + (1 | panchayat)
##    Data: data
##
## REML criterion at convergence: 2989.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.5084 -0.6397  0.0021  0.6152  3.3527
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  panchayat (Intercept) 0.2502   0.5002
##  Residual              1.7010   1.3042
## Number of obs: 877, groups:  panchayat, 3
##
## Fixed effects:
##               Estimate Std. Error t value
```

```
## (Intercept)  4.510736   0.399010  11.305
## age          0.138626   0.002468  56.168
## sexfemale   -0.380090   0.089220  -4.260
## lit          1.037553   0.282254   3.676
## alive       -0.112921   0.035360  -3.194
## mage         0.071060   0.012485   5.692
## died         0.089830   0.056580   1.588
##
## Correlation of Fixed Effects:
##           (Intr) age    sexfml lit    alive  mage
## age       -0.120
## sexfemale -0.104  0.031
## lit       -0.158 -0.008 -0.075
## alive      0.361 -0.016  0.055 -0.061
## mage      -0.591 -0.125 -0.032  0.111 -0.778
## died      -0.075 -0.006 -0.101  0.070 -0.523  0.225

## refitting model(s) with ML (instead of REML)
```


Posterior Predictive Check
Model−predicted lines should resemble observed data line
Density
wt
Observed data
Model−predicted data
Linearity
Reference line should be flat and horizontal
Residuals
Fitted values
Homogeneity of Variance
Reference line should be flat and horizontal
√|Std. residuals|
Fitted values
Influential Observations
Points should be inside the contour lines
Std. Residuals
Leverage (hii)
0.9
344
348
716
Collinearity
High collinearity (VIF) may inflate parameter uncertainty
Variance Inflation Factor (VIF, log−scaled)
age
alive
died
lit
mage
sex
Low (< 5)
Normality of Residuals
Dots should fall along the line
Sample Quantiles
Standard Normal Distribution Quantiles
Normality of Random Effects (panchayat)
Dots should be plotted along the line
RE Quantiles
Theoretical Quantiles

```
## Warning: Removed 123 rows containing missing values (`geom_point()`).
```

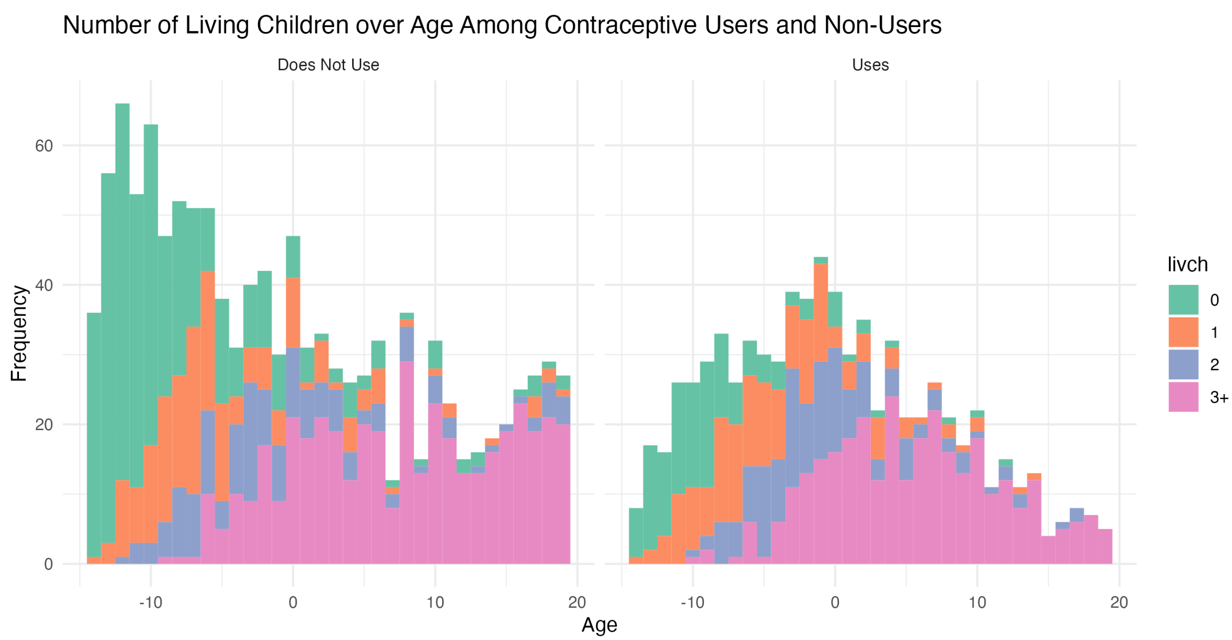


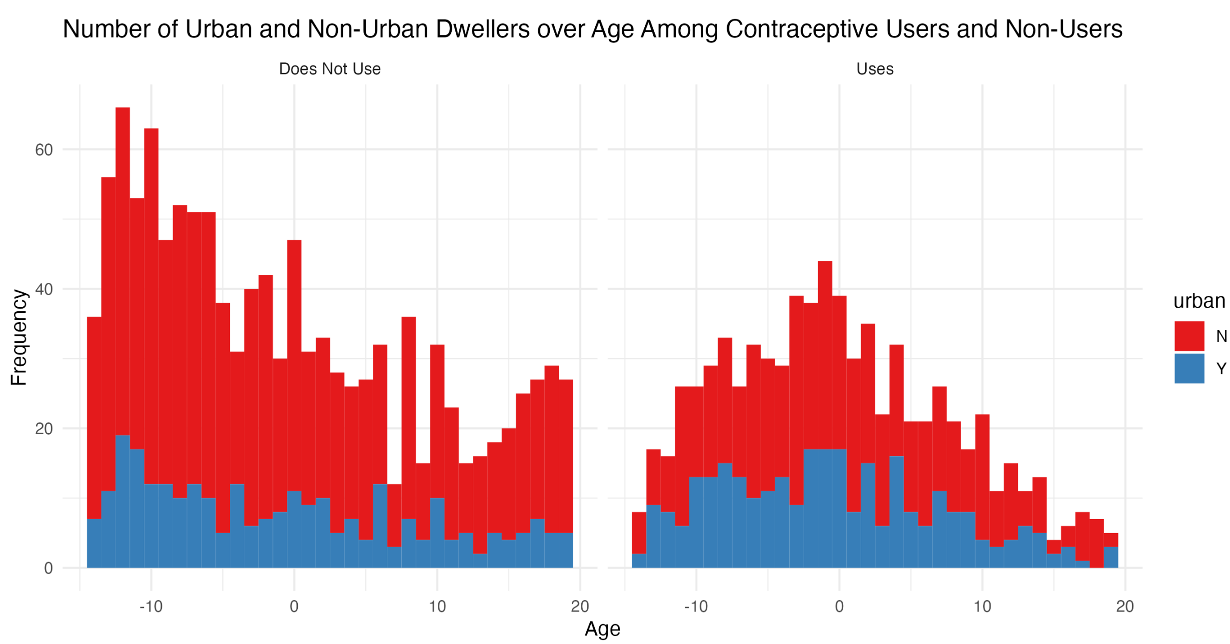
The paper by West et al describes a double-masked, randomized, placebo-controlled community trial assessing the impact of high-potency vitamin A supplementation on the growth of preschool-aged children in Nepal. This design involves a direct intervention (vitamin A supplementation) and a control group. The study employed chi-square tests for categorical variables and analysis of variance for continuous variables to evaluate baseline group differences. Growth increments were compared using linear regression, adjusted for age, baseline values, and sex. The regression analysis also considered arm circumference as an effect modifier. The analysis was stratified by initial arm circumference to account for children who were wasted and not wasted at the outset.

Our approach involved using a linear mixed effects model with the formula `wt ~ age + sex + lit + died + alive + mage + (1|panchayat)`, applied to the `faraway::nepali` dataset. This model focused Focuses on exploring the relationship between various predictors and child weight. We included a random effect for panchayat to account for variability across different panchayats.

The paper's method is based on a controlled trial with a specific intervention (vitamin A supplementation), whereas our approach is more observational, exploring relationships without a specific intervention. The paper's study involves randomization and control groups whereas our approach does not involve these elements. The paper focuses on growth impacts (with measurements taken periodically) in the context of a vitamin A intervention, whereas our model looks at weight in relation to various predictors.

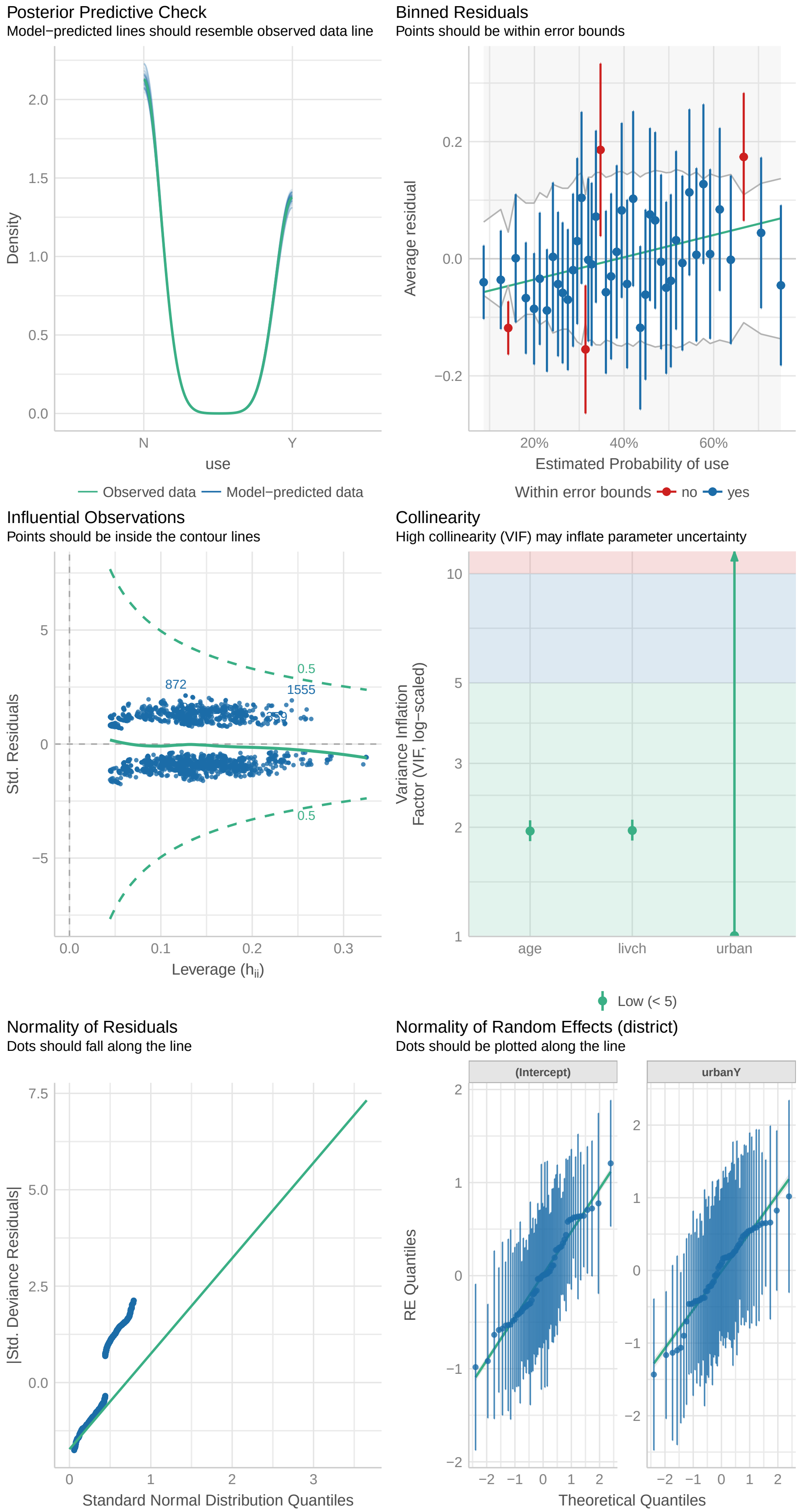
2 We fit a linear model with a logit link function to analyze a binary outcome. Specifically, we used the `lme4::glmer` function to construct two models incorporating random effects. The first model included a random intercept based on the urban or rural status of districts. The second model extended this by adding a random slope for the urban classification of districts. Upon comparing these models using a Likelihood Ratio Test (LRT), we found that the model with both random intercept and slope for the urban variable `(urban | district)` demonstrated superior fit. This was evidenced by lower values in Akaike Information Criterion (AIC), log-likelihood, and deviance, although it had a slightly higher Bayesian Information Criterion (BIC). The Chi-square test revealed a significant improvement with this model (p-value = 0.0006759), leading us to adopt the formula `use ~ age + livch + age + urban + (urban | district)` for our final model. We then assessed the model's fit using the `performance::check_model` function, which indicated a generally good fit but highlighted some unusual patterns in the residuals. To further investigate, we applied the dHARMa package, leading to the residual plot provided below. This plot confirms that the residuals of our chosen model are appropriately distributed. Regarding the model coefficients, the `age` coefficient is -0.026518, indicating that with each additional year of age, the log odds of using contraception decreases. The coefficients for the number of living children (`livch`) are all positive and exceed 1. This suggests that having one, two, or more than three living children significantly increases the likelihood of contraception use compared to having no living children. Lastly, the coefficient for `urban` (0.815146) implies that residing in an urban area, as opposed to a non-urban one, increases the odds of contraception use.



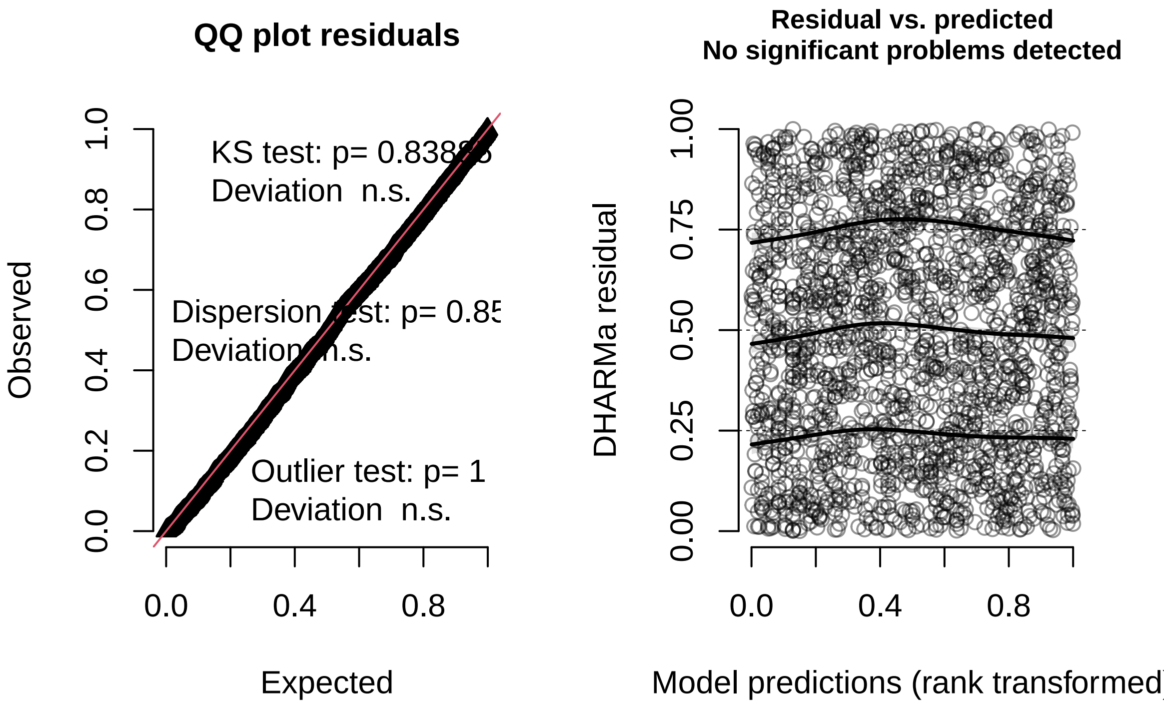


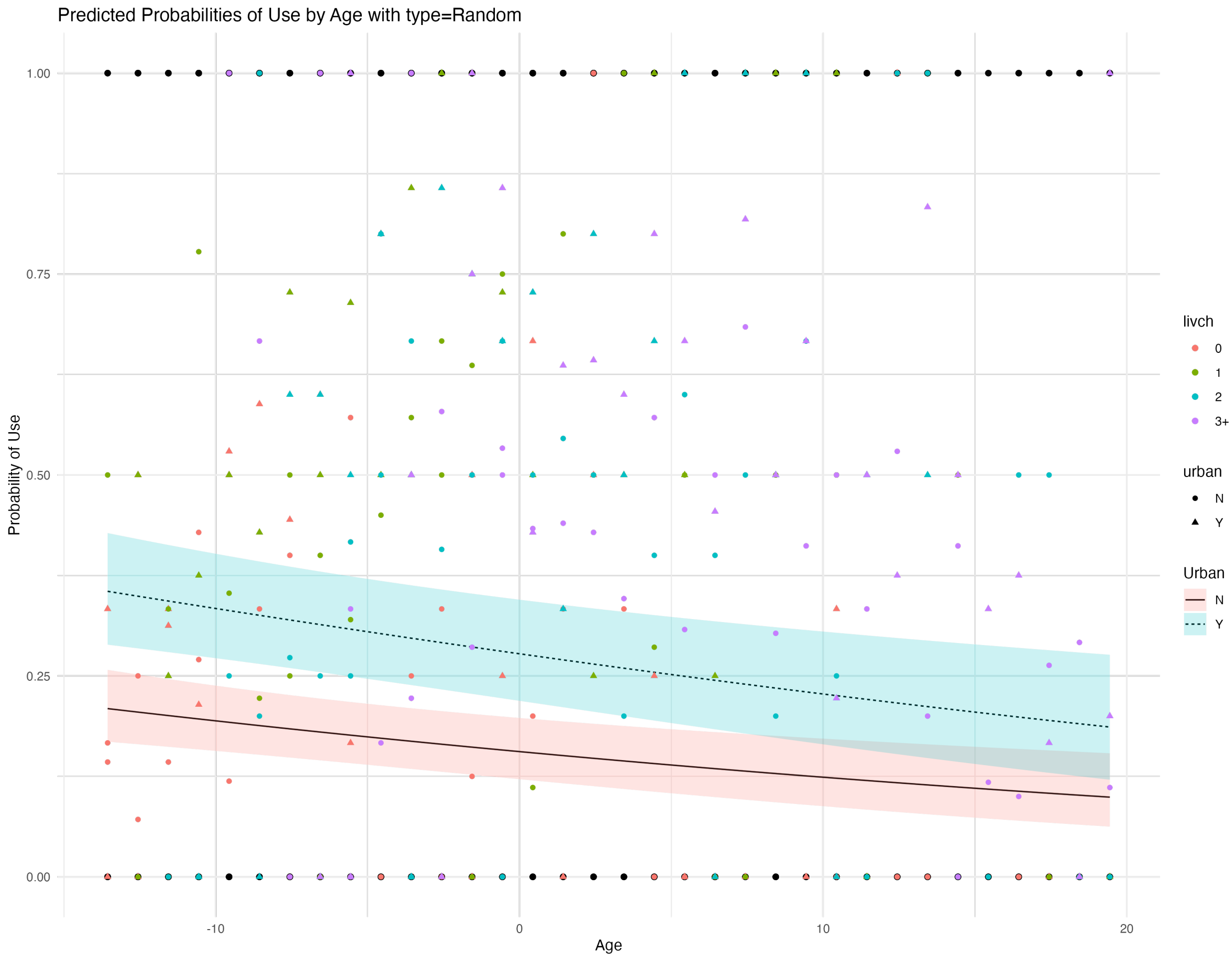

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: use ~ age + livch + age + urban + (urban | district)
##    Data: Contraception
##
##      AIC      BIC   logLik deviance df.resid
##   2417.0   2467.1  -1199.5   2399.0     1925
```

```
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.9127 -0.7456 -0.4933  0.9335  2.9272
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  district (Intercept) 0.3811   0.6173
##           urbanY      0.6418   0.8011   -0.80
## Number of obs: 1934, groups:  district, 60
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.711708   0.159621 -10.724  < 2e-16 ***
## age         -0.026518   0.008001  -3.314 0.000918 ***
## livch1       1.125641   0.159900   7.040 1.93e-12 ***
## livch2       1.368212   0.176849   7.737 1.02e-14 ***
## livch3+      1.354720   0.182410   7.427 1.11e-13 ***
## urbanY       0.815146   0.169719   4.803 1.56e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##         (Intr) age    livch1 livch2 lvch3+
## age      0.422
## livch1  -0.550 -0.212
## livch2  -0.590 -0.380  0.487
## livch3+ -0.701 -0.675  0.538  0.617
## urbanY  -0.473 -0.035  0.042  0.066  0.062
```

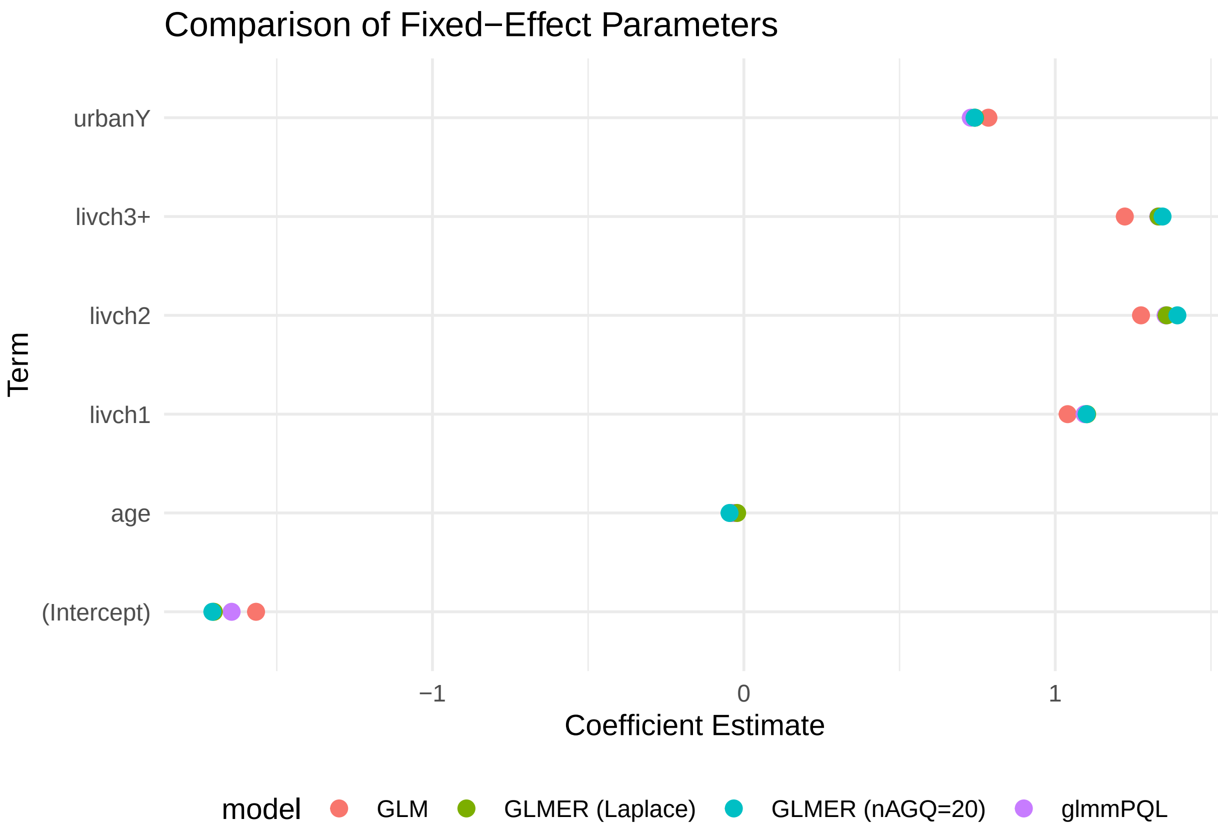

Posterior Predictive Check
Model−predicted lines should resemble observed data line
Density
use
N
Y
Observed data
Model−predicted data
Binned Residuals
Points should be within error bounds
Average residual
Estimated Probability of use
Within error bounds
no
yes
Influential Observations
Points should be inside the contour lines
Std. Residuals
Leverage (hii)
872
1555
0.5
Collinearity
High collinearity (VIF) may inflate parameter uncertainty
Variance Inflation Factor (VIF, log−scaled)
age
livch
urban
Low (< 5)
Normality of Residuals
Dots should fall along the line
|Std. Deviance Residuals|
Standard Normal Distribution Quantiles
Normality of Random Effects (district)
Dots should be plotted along the line
(Intercept)
urbanY
RE Quantiles
Theoretical Quantiles

DHARMa residual





Predicted Probabilities of Use by Age with type=Random
1.00
0.75
0.50
0.25
0.00
Probability of Use
-10
0
10
20
Age
livch
0
1
2
3+
urban
N
Y
Urban
N
Y



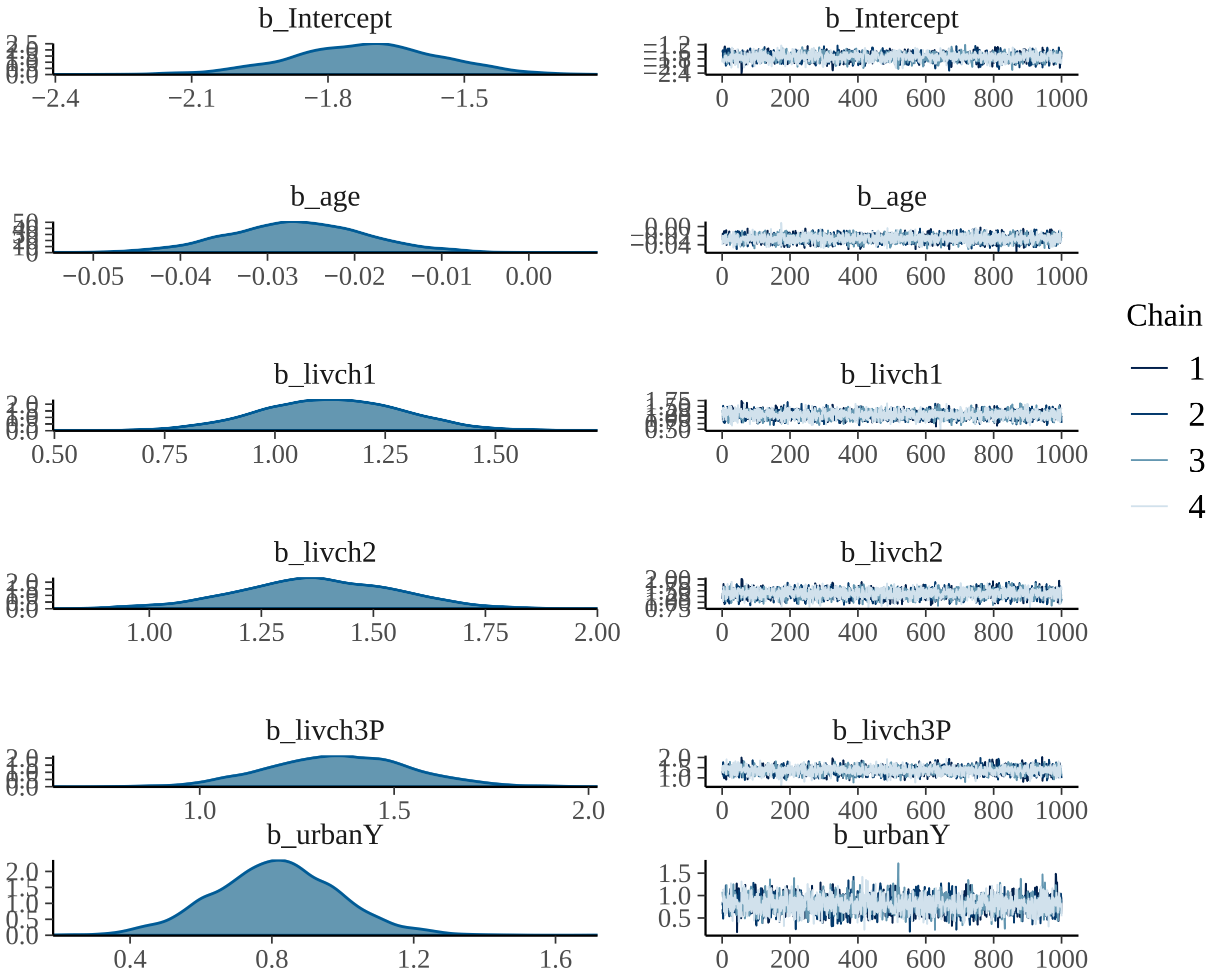
The research paper by Ng et al. focuses on estimating generalized linear mixed models, specifically those with binary outcomes. The main challenge addressed in the paper is the difficulty in obtaining an analytical solution for the likelihood of a discrete response. This issue is significant because it can lead to bias in the results obtained through marginal and penalized quasi-likelihood methods. To explore this problem, the authors the a dataset referred to as "BANG." They applied several modeling techniques to this dataset, including Second Order PQL (PQL_2), SML (Sequential Monte Carlo), EM_Laplace2, MCMC (Markov Chain Monte Carlo), as well as numerical quadrature methods implemented through Proc_NLMIXED and GLLAMM. They used the same model formula as our analysis. This approach allows for the examination of variations in variances across different districts, particularly between rural and urban areas. Secondly, it provides a basis for comparison with the `lme4::glmer` function used in our analysis, which employs a Laplace approximation method. When comparing the coefficient estimates from our model to those from the EM_Laplace2 model used in the Ng et al. paper, we find that the values are somewhat similar. The differences in the fixed effects parameters across these two models are minimal, with discrepancies of at most 10^{-2}.

3

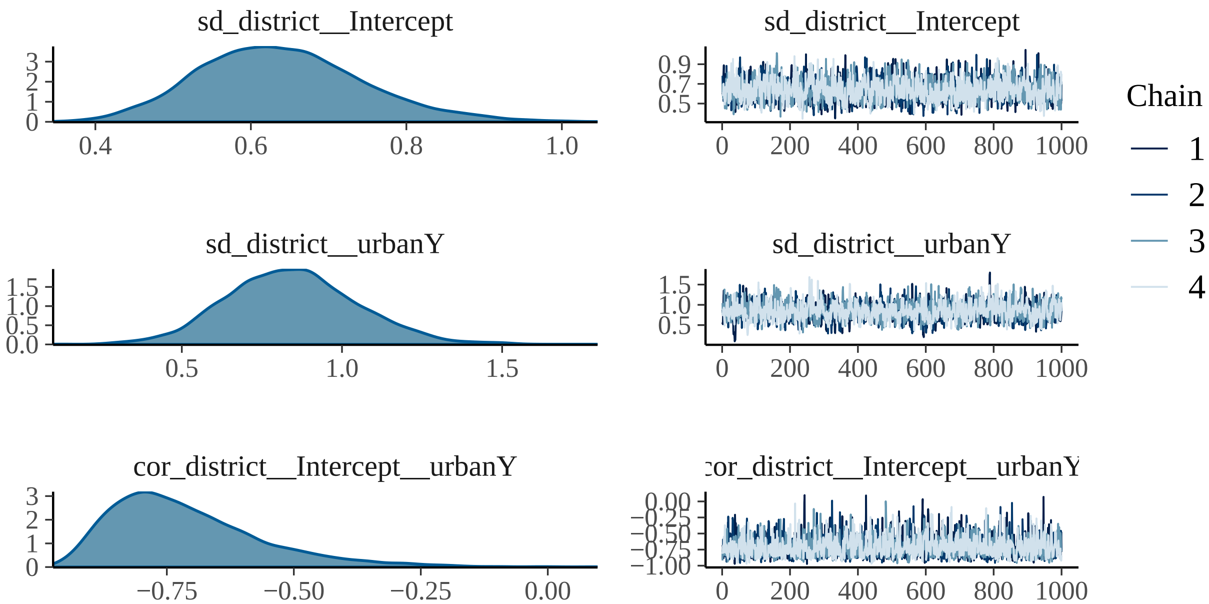
We fit the same mode formula as in q2 with the Contraception data set using two bayesian methods: `brms::brm` and `rstanarm::stan_glmer`. All parameter estimates has sufficently large effective sample sizae and $\hat{R}$ was 1 for all covariates. As you an see in the figures below, both models converged and the chains mixed well. The brm model overestimated on the order of 10^-3 on average across all fixed effect parameter estimates.

```
##  Family: binomial
##   Links: mu = logit
## Formula: use | trials(n_trials) ~ age + livch + urban + (urban | district)
##    Data: Contraception (Number of observations: 1934)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Group-Level Effects:
## ~district (Number of levels: 60)
##                       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
## sd(Intercept)             0.64      0.11     0.45     0.87 1.00     1673
## sd(urbanY)                0.85      0.21     0.46     1.27 1.00     1446
## cor(Intercept,urbanY)    -0.71      0.15    -0.92    -0.34 1.00     1808
##                       Tail_ESS
## sd(Intercept)             2566
## sd(urbanY)                2155
## cor(Intercept,urbanY)     2503
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept    -1.72      0.16    -2.04    -1.41 1.00     2163     2507
## age          -0.03      0.01    -0.04    -0.01 1.00     3964     3118
## livch1        1.13      0.16     0.81     1.45 1.00     3813     3217
## livch2        1.37      0.18     1.01     1.71 1.00     3562     3159
## livch3P       1.36      0.18     1.01     1.72 1.00     2837     2555
## urbanY        0.81      0.18     0.46     1.17 1.00     2598     2767
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
```

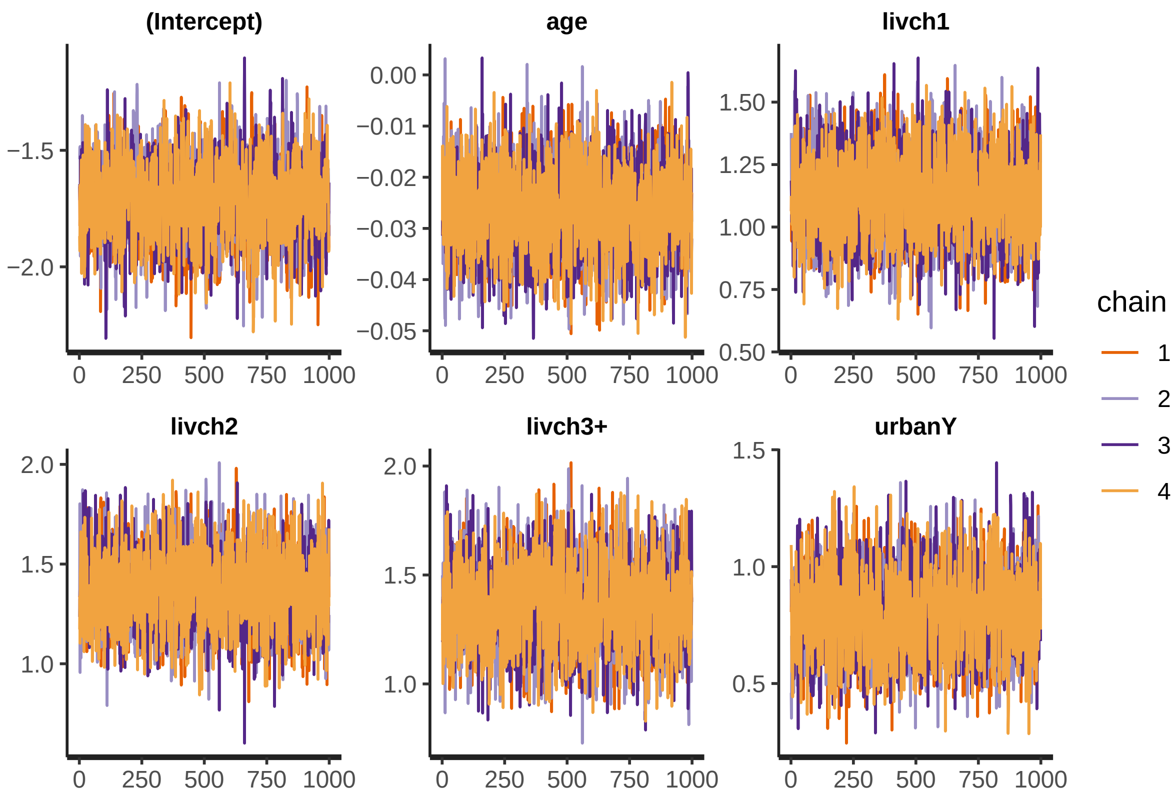
```
## scale reduction factor on split chains (at convergence, Rhat = 1).
##                    mean      se_mean          sd        2.5%         10%
## (Intercept) -1.71142387 0.0034393395 0.161404189 -2.03507306 -1.91522842
## age         -0.02645698 0.0001368276 0.008165125 -0.04319387 -0.03690346
## livch1       1.12805203 0.0027687448 0.163228892  0.82131787  0.91746826
## livch2       1.37058776 0.0031253230 0.180697776  1.02231475  1.14222655
## livch3+      1.35781374 0.0037554624 0.188950266  0.99322479  1.11611562
## urbanY       0.79900113 0.0034389834 0.170030314  0.47402214  0.58081247
##                     25%         50%         75%         90%       97.5%
## (Intercept) -1.81760090 -1.70948143 -1.60550983 -1.50806120 -1.39582024
## age         -0.03190429 -0.02654181 -0.02086206 -0.01601351 -0.01103207
## livch1       1.01395024  1.12741761  1.23948369  1.34240045  1.45088977
## livch2       1.24721695  1.36766364  1.49191328  1.59634842  1.74717994
## livch3+      1.22509768  1.35440262  1.48638970  1.60140384  1.73935733
## urbanY       0.68258741  0.79824654  0.90780124  1.01616221  1.14216252
##                n_eff     Rhat
## (Intercept) 2202.315 1.002019
## age         3561.051 1.000132
## livch1      3475.587 1.000220
## livch2      3342.842 1.001430
## livch3+     2531.443 1.001020
## urbanY      2444.514 1.000059
```

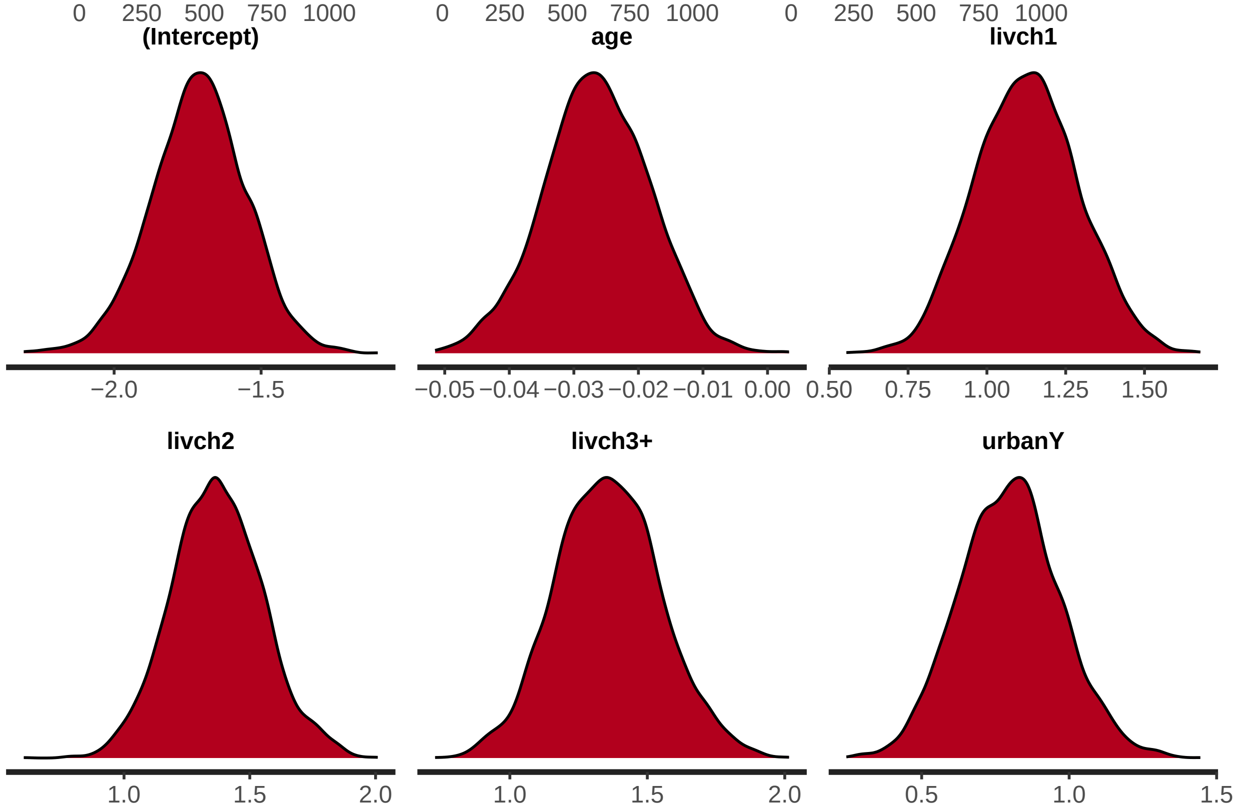


b_Intercept
−2.4
−2.1
−1.8
−1.5
b_Intercept
0
200
400
600
800
1000
b_age
−0.05
−0.04
−0.03
−0.02
−0.01
0.00
b_age
b_livch1
0.50
0.75
1.00
1.25
1.50
b_livch1
b_livch2
1.00
1.25
1.50
1.75
2.00
b_livch2
b_livch3P
1.0
1.5
2.0
b_livch3P
b_urbanY
0.0
0.5
1.0
1.5
2.0
0.4
0.8
1.2
1.6
b_urbanY
Chain
1
2
3
4

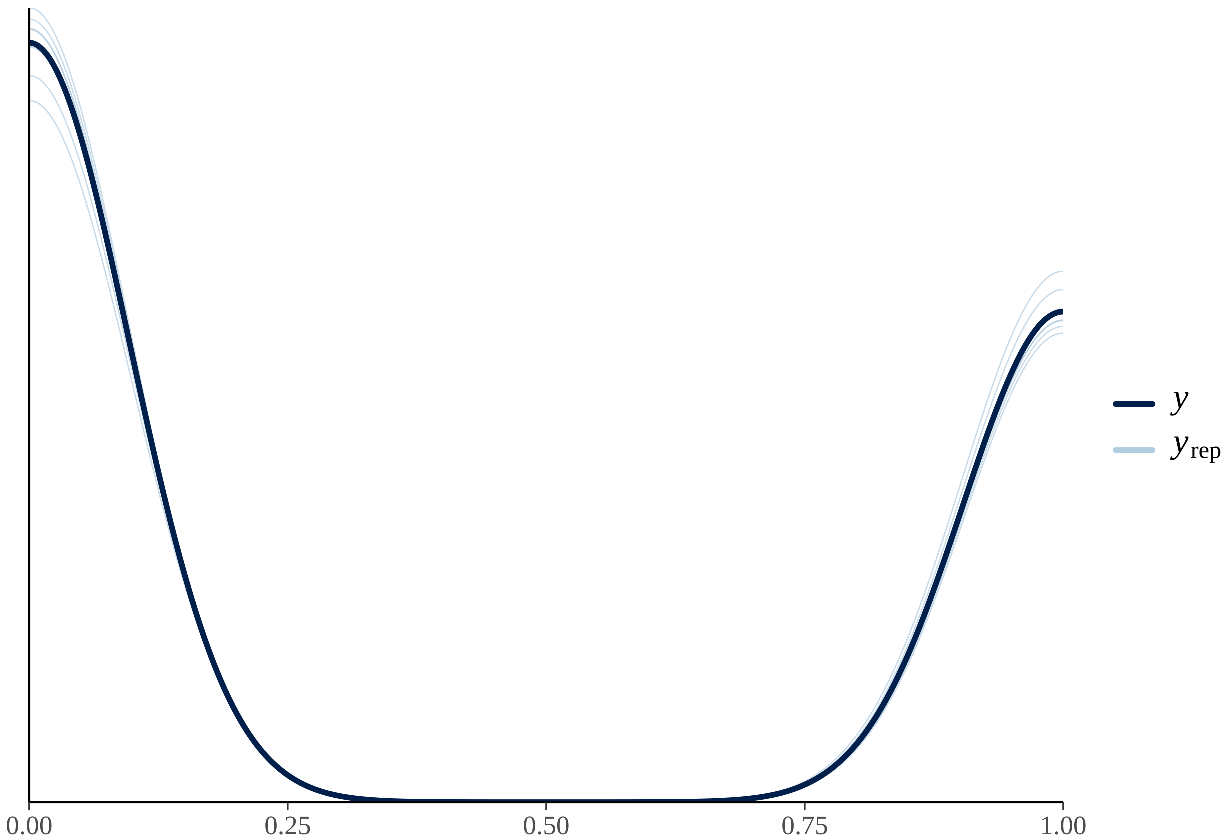


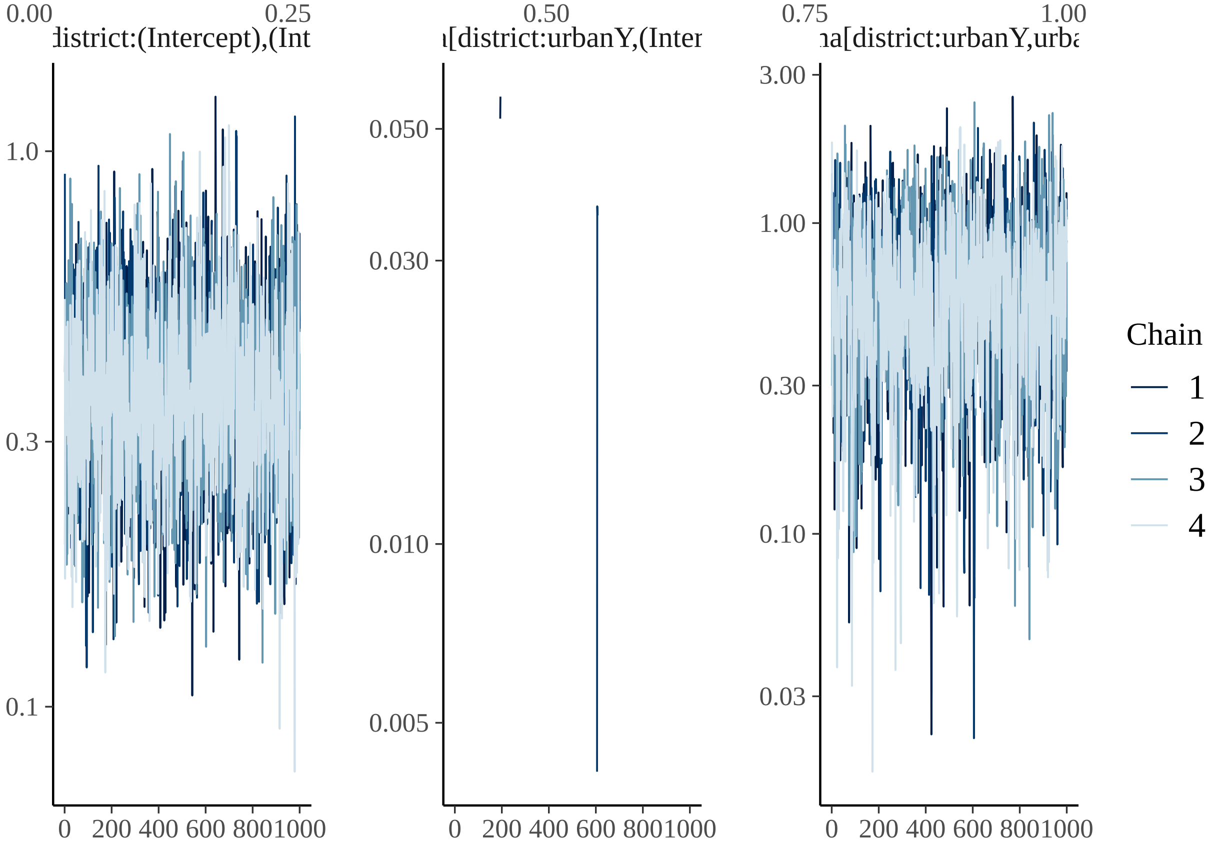
sd_district__Intercept
0
1
2
3
0.4
0.6
0.8
1.0
sd_district__Intercept
0.5
0.7
0.9
0
200
400
600
800
1000
sd_district__urbanY
0.0
0.5
1.0
1.5
sd_district__urbanY
cor_district__Intercept__urbanY
−0.75
−0.50
−0.25
0.00
Chain
1
2
3
4





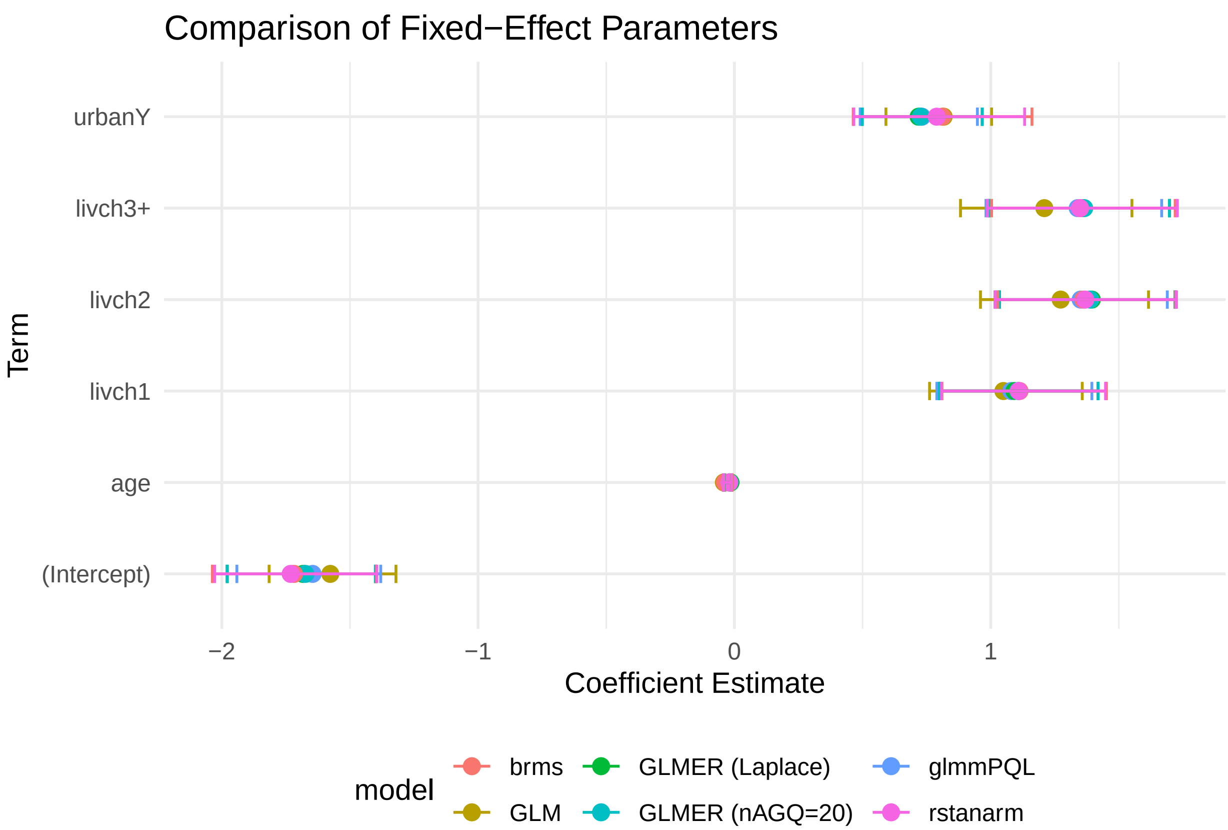

```
## Error : `check_model()` not implemented for models of class `glmerMod` yet.
```






```
##                                    Parameter      Rhat      ESS        MCSE
```

```
## 126 Sigma[district:(Intercept),(Intercept)] 1.000843 1772.496 0.003253456
## 127       Sigma[district:urbanY,(Intercept)] 1.000106 1359.062 0.004717159
## 128            Sigma[district:urbanY,urbanY] 1.000543 1378.297 0.008990404
```



4

Metric	nAGQ Value	Beta[1] Value	Value
Bias	-2	-2	-0.0055111
Variance	-2	-2	0.0110776
Scaled RMSE	-2	-2	0.2097352
Coverage	-2	-2	0.9400000
Bias	-2	2	0.0088797
Variance	-2	2	0.0059406
Scaled RMSE	-2	2	0.1544031
Coverage	-2	2	0.9300000
Bias	-1	-2	-0.0156817
Variance	-1	-2	0.0087174
Scaled RMSE	-1	-2	0.1884269

Metric	nAGQ Value	Beta[1] Value	Value
Coverage	-1	-2	0.9600000
Bias	-1	2	0.0002725
Variance	-1	2	0.0001771
Scaled RMSE	-1	2	0.0264864
Coverage	-1	2	0.9300000
Bias	1	-2	-0.0126013
Variance	1	-2	0.0120597
Scaled RMSE	1	-2	0.2199804
Coverage	1	-2	0.9400000
Bias	1	2	-0.0005454
Variance	1	2	0.0001882
Scaled RMSE	1	2	0.0273244
Coverage	1	2	0.9200000
Bias	2	-2	-0.0202737
Variance	2	-2	0.0132863
Scaled RMSE	2	-2	0.2329329
Coverage	2	-2	0.9500000
Bias	2	2	0.0000443
Variance	2	2	0.0001574
Scaled RMSE	2	2	0.0249670
Coverage	2	2	0.9400000