Investigating birth control uptake across 60 regions using bayesian logistic regression (and STAN)

Anonymous

09/04/2021. Length: 8 pages excluding title page, appendix and citations.

Introduction

A pharmaceutical company that markets a birth control drug is looking to get a better understanding of their potential customer base. In this project this will be done by investigating a range of demographic variables of those who take and do not take birth control. The effect and type of effect of these demographic variables will be investigated in order to predict the chance of certain demographics purchasing birth control in the future.

First, exploratory data analysis will be done to understand the provided data and from this a type of model will be selected. After this, the model will be built and it's conclusions and predictive power will be investigated.

Methods

Exploratory Data Analysis

"birthControl" is the outcome variable with two levels, 0 and 1, meaning a logistic regression will be used.

There are 60 of "region", and "homeStyle" has two levels - "urban" and "rural". "children" gives the amount of children the subject has and their age and wealth are standardised continuous variables.

Missing Data

A missing data check shows there is no missing data in this dataset.

Exploring the variables

Figure 1 shows the distribution of region split by takers and non-takers of birth control. It can be seen that region 3, 11 and 49 have only one type of birth control takers, for example in region 3 all of the observations. This means for the model random effects must be used as fixed effects requires data uptakers and non-uptakers of birth control in each region.

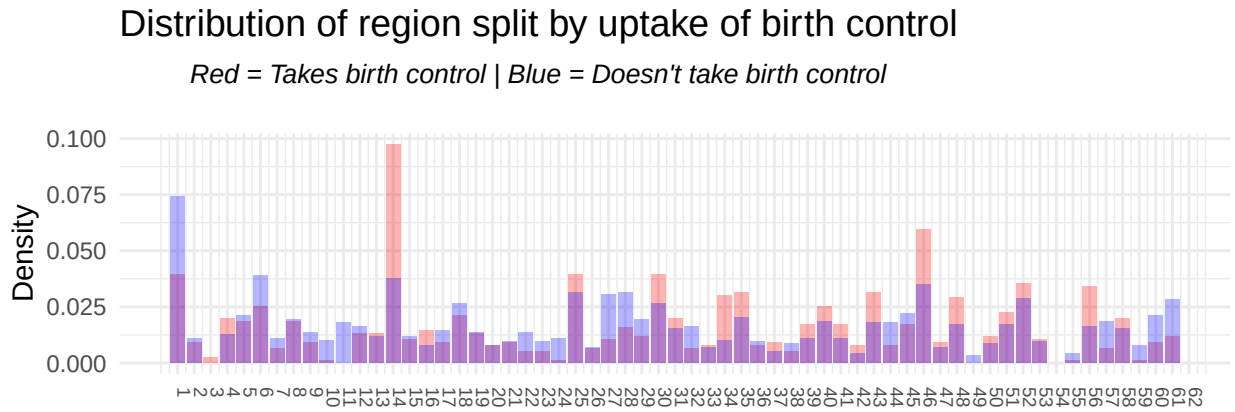

Figure 1: Distribution of region split by uptake of birth control. This highlights regions 3, 11 and 49 having just one unique value for the birth control variable, motivating the use of random effects

It's also noted that region 54 contains no data and this can be troublesome in the model build process, for that reason the data is recorded so regions 55-61 inclusive are given the value of region 1 below their original value (e.g region 55 becomes region 54).

Figure 2 shows the distribution of ages that take birth control and don't take birth control. Those that take birth control increases as women reach average age. As they get older this relationship flips as older women take more birth control.

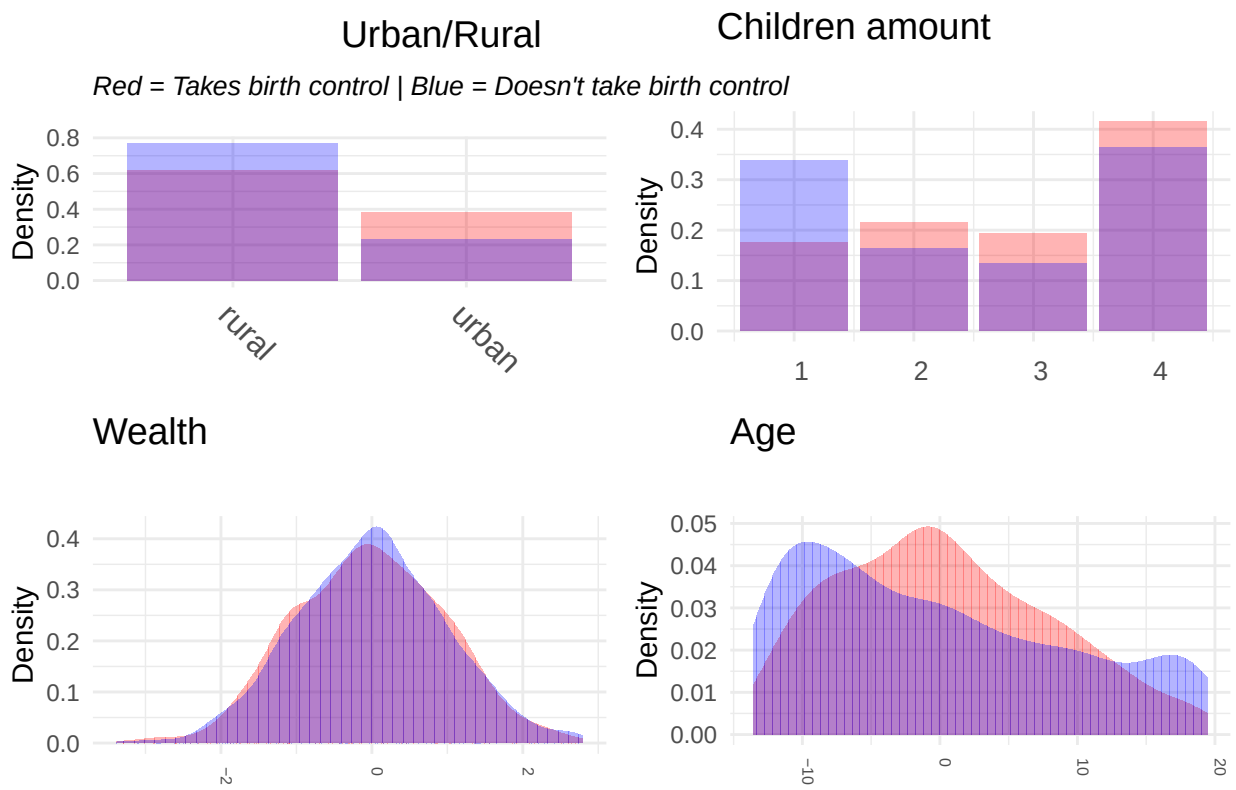

Figure 2: Distributions of the independent varialbles split by yes/no uptake of birth control.

Figure 2 shows a plot for wealth. There's less variance in those that don't take birth control and both are normally distributed.

Figure 2 shows the distribution of the amount of children split by uptake of birth control. Those with 1 child take less birth control than those who do. As women have 2 children or more this flips and there are more birth control takers than none-takers.

Figure 2 shows that those in urban areas proportionally take more birth control, and the reverse is true for those in live in rural areas.

Figure 3 shows the distribution of birth control takers. In the data around 39% of observations take birth control.

Between the independent variables (except region), all combination of groups have values, allowing for interactions if necessary and they can be included as fixed.

Relationships

It's already been seen that those who live in urban have a higher rate of birth control uptake. Figure 4 looks at how the relationship of the independent variables with birth control by plotting a scatter plot and displaying a regression line. It's seen that wealthier observations have less uptake of birth control, but this relationship is very weak. Higher age and higher number of children is associated with more uptake of birth control, with the number of children showing the strongest relationship.

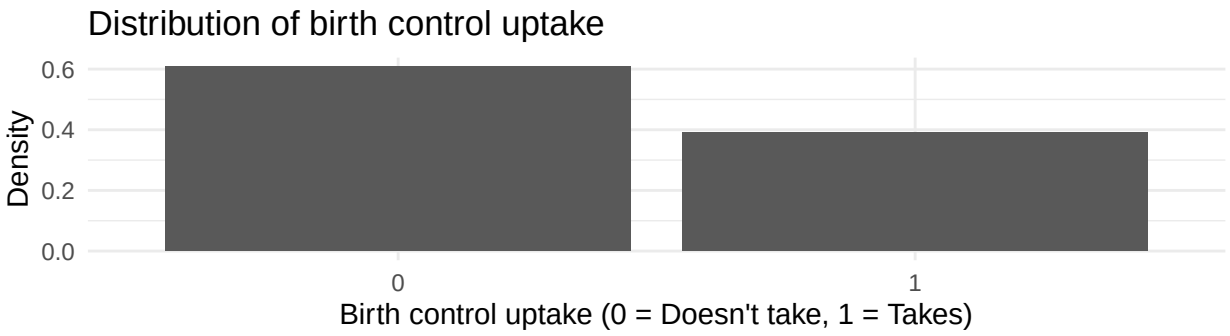

Figure 3: Distribution of birth control uptake.

The bottom right quadrant of this figure shows red regression lines for each region for the relationship between the number of children and birth control uptake. It can be seen that there are different relationships between these two variables within the region which highlights that the different regions have different relationships between children and birth control, which will later motivate using a regression which allows for these regional differences.

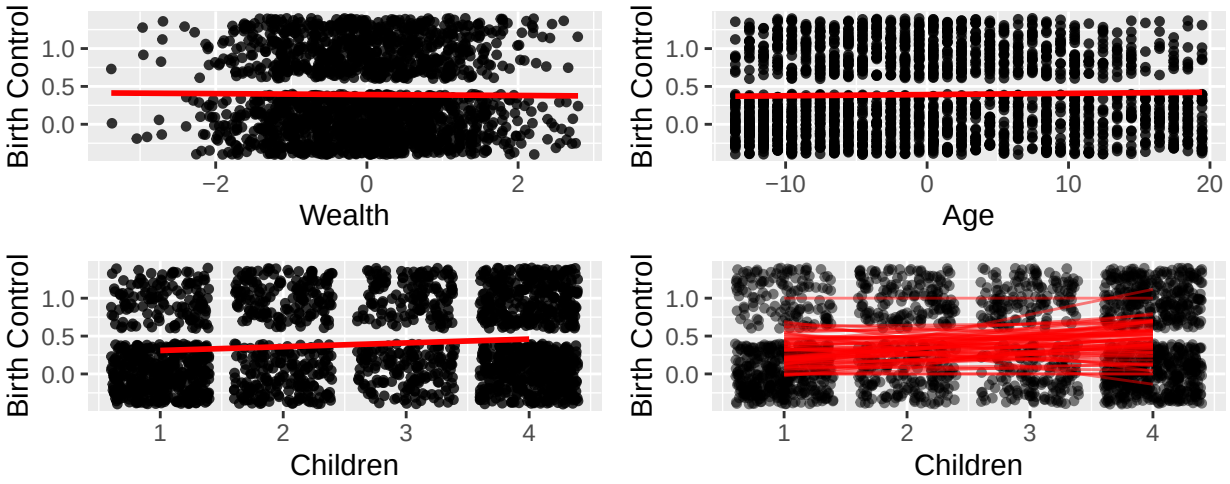

Figure 4: An exploration of how the independent variables relate with birth control. The bottom right plot shows the relationship between the amount of children and birth control, with the red regression lines representing this relatonship within each region.

Table 1 shows the correlation between the continuous independent variables. The only variables that show a very high correlation are the age and the amount of children variable. Having both of these variables in the model will not give us much addition information, and since, from Figure 4, the amount of children has the strongest relationship with birth control uptake (higher gradient), it is decided that age will not be used going forward into the analysis.

Table 1: Correleation values between the independent variables

	children	age	wealth
children	1.00	0.69	-0.01
age	0.69	1.00	-0.02
wealth	-0.01	-0.02	1.00

Exploring Interactions

In Figure 5 the left hand plot shows a scatter plot of birth control uptake against wealth, with lines representing how the relationship between these to variables vary with the amount of children. It can be seen that the relationship is different between the number of children; for the data with 3 or 4 children, wealthier subjects have less birth control uptake. For those with 1 or 2 children, wealthier subjects have more birth control uptake. This suggests an interaction between number of children and wealth should be used.

The middle plot shows uptake of birth control against the number of children, with the lines representing the relationship between these variables in the rural and urban groups. The relationship between these two groups is the same, having straight lines with the same gradient, so an interaction term between children and amount and living location of the subject is not considered.

The right hand plot shows birth control uptake against wealth with the lines representing the relationship between these variables split by rural and urban groups. These lines shows rural/bran groups have different relationships within the birth control and wealth relationship, suggesting an interacting term of wealth and location should be used.

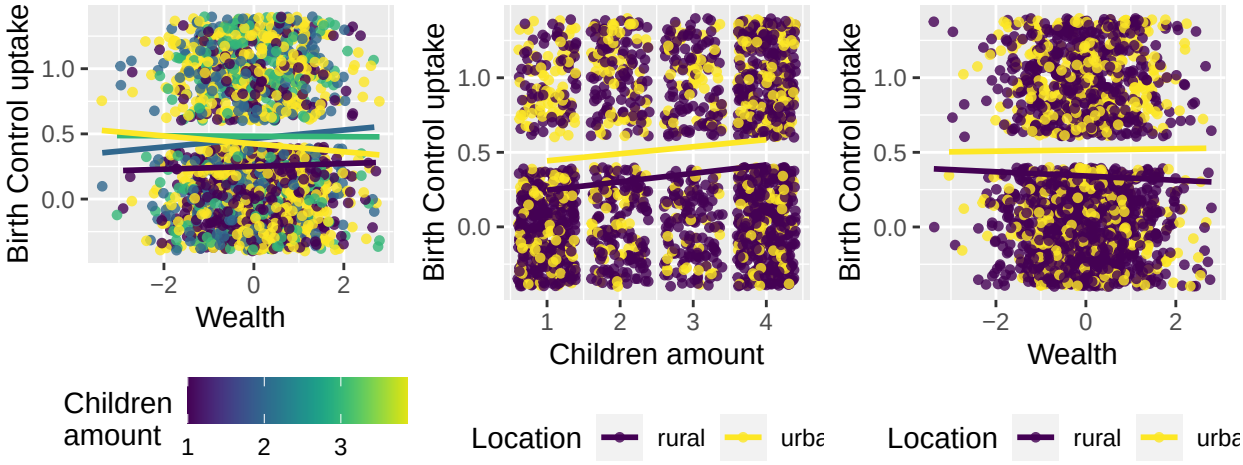

Figure 5: An exploration of how the independent variables relate with each other. The plot on the left shows birth control uptake against wealth, with the lines representing how the relationship of these two variables vary dependending on the amount of children the observation has. The right plot shows birth control uptake against the number of children with the lines showing how the relationship between these two variables split by urban/rural living.

Final model method

This project will include two regressions. The first will be a Bayesian logistic regression that has a random effect for each region the observation belongs to. This random effect term will come from an uninformative prior.

The second regression will be a hierarchical model regression. Unlike the first regression, this does not assume that the effect of our independent variables are not the same across the regions. It allows for regional level variation of the effect sizes of the coefficients of the independent variables by drawing from a population level distribution. Allowing for this regional level variation makes sense since it should not be expected that the independent variables have the same effect across all regions. For example, for one region maybe having more children has a different effect on the uptake of birth control than another region.

Prior distributions are assigned to each of the regression parameters with independent standard deviations for each region. An overall estimate of the effect size for a random region be done by independently sampling from their prior distributions.

These prior distributions are from a Normal(0,1) distribution and Andrew Gelmen described this as a **"Generic weakly informative prior"**.

This has a similar method to running a separate model for each region (heterogeneous model) however with the hierarchical model there is dependence between individual region level parameters. This means there is less variance in the hierarchical model than the heterogeneous one since information is pooled across all the different regions. Also, parameters estimates are closer to the overall mean than the heterogeneous model parameters.

The priors used in the hierarchical model are normal distributions.

Both regressions with have the following independent variables:

- **homeStyle** : Whether the observation lives in rural or urban setting.
- **Wealth** : How wealthy the observation is - this variable is standaridsed.
- **Children** : Indicator variable with 0 = having 1 or 2 children, 1 = having 3 or 4 children
- **Wealth x homeStyle** : Interaction terms between wealth and homeStyle.
- **Wealth x Children** : Interaction term between wealth and children.

Results

Regression with uninformative prior

For the uninformative prior regression, the results' independent variables have rhats and n_effs that show good convergence however the intercept has n_eff below 100 meaning it has convergence issues.

The trace plots also convergence in all variable except the intercept. This is a problem, and the regression likely needs more interactions to mend this. The intercept failed to converge for both 2000 and 3000 iterations. Since this regression is not the main focus of the project, the results will be left as is. Seemingly an uninformed prior means the intercept does not converge.

Hierarchical model

Model diagnostic metrics are given in the appendix.

Table 2 shows the parameter estimates for the hierarchical model. The estimates of Rhat close to 1 and n_eff > 100 which shows good convergence of the chains.

Table 2: Regression results from hierarchical model. The means column gives the effect of each independent variable

	mean	se_mean	sd	2.5%	50%	97.5%	n_eff	Rhat
Intercept	-0.25	0.01	0.35	-1.00	-0.24	0.45	933.63	1.00
Urban/Rural	-0.86	0.02	0.78	-2.48	-0.87	0.65	1672.91	1.00
Wealth	0.12	0.01	0.13	-0.15	0.12	0.36	110.61	1.03
Children	0.06	0.05	1.49	-2.92	0.07	3.08	1087.07	1.00
Wealth*Children	-0.18	0.01	0.24	-0.65	-0.19	0.28	565.98	1.00
Wealth*Urban/Rural	-0.10	0.01	0.20	-0.49	-0.10	0.27	174.98	1.02

Figure 6 shows the trace plots for the regression with all the parameters converging which is wanted.

Convergence was also seen in the non-hierarchical model, the first regression, when using a normal prior for the random effects of the region values. The results of that regression will not be discussed/printed here, however it is noted that in this project the use of an informative prior allows for convergence of the intercept and we can see the same convergence in the hierarchical model which uses normal informative priors.

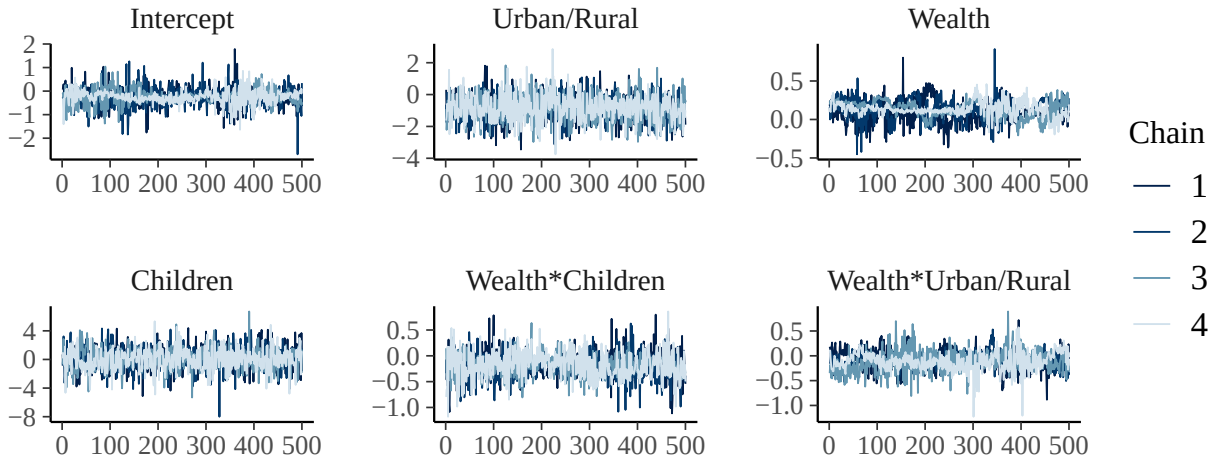

Figure 6: Trace plots of the sampled parameters for the hierarchical model

Figure 7 shows the effect of each independent variable on birth control uptake. No variable is deemed significant at the 5% level as their confidence intervals all include 0.

Being wealthier means a woman is more likely to take birth control, as the parameters lie above 0. This relationship also applies to having 3 or 4 children (compared to 1 or 2).

Wealthier women that also have 3/4 children decreases the likelihood of uptake of birth control - to put simply, if you have 1/2 children, the wealthier you are the more likely to take birth control The wealthier you are when you have 3/4 children, the less likely you are to take birth control.

Wealthier women that live in rural areas also decreases this likelihood. So the wealthier you are in an urban area, you're more likely to take birth control. The wealthier you are in a rural area, you're less likely to take birth control.

Rural women are less likely to take birth control.

Using this information, the company may want to focus their product in urban areas, especially wealthier women and those that have 1/2 children since although having 3/4 children has a positive effect on the uptake of birth control, this is offset by the interaction of wealth*children by a larger amount. It must be started that none of these variables are deemed significant at the 5% level so these relationships may exist purely by chance (i.e we cannot reject the null hypothesis that the relationships have no relationship with birth control uptake).

Prior check

As mentioned before, Gelman suggests the use of a weakly informative prior of *N(0,1)*. This means when the prior is compared with the posterior, because of the weakness, a big difference isn't expected.

Figure 8 shows the prior and posterior distributions with the red line showing the likelihood. In general it can be seen the estimates of the model follow the prior distribution as expected with such a weakly informative prior ie it allows for a high range of values in the context of a regression.

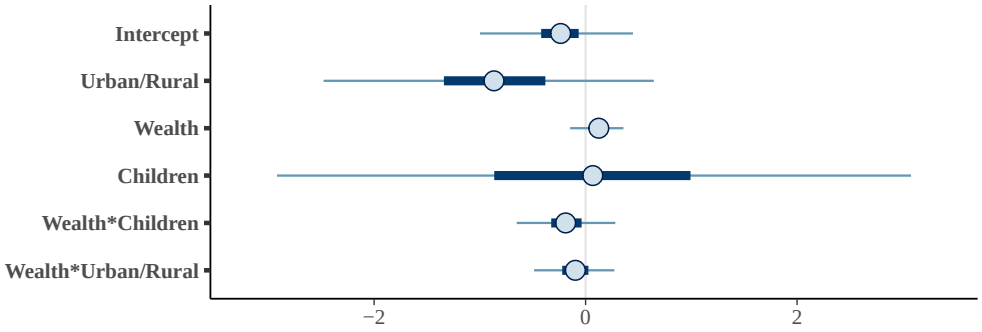

Figure 7: Posterior credible interval summaries for birth control uptake for the hierarchical model, showing the effect of each variable on uptake of birth control. Point gives the posterior median, bar gives the interval with quantiles at 0.25 and 0.75, and the line gives the quantiles at 0.025 and 0.975.

The data is allowed influence on the posterior distribution as seen in the plot for Urban/Rural (2nd plot). The prior is pulled to the left of the prior distribution, with it's median closer to the likelihood - demonstrating the ability for the data to have a big influence on the prior distribution. This also happens for wealth and the intercept.

The priors do follow the distribution of the prior distributions for wealth*children , which is a cause for concern. This concern can be addressed by opting for a more informative prior.

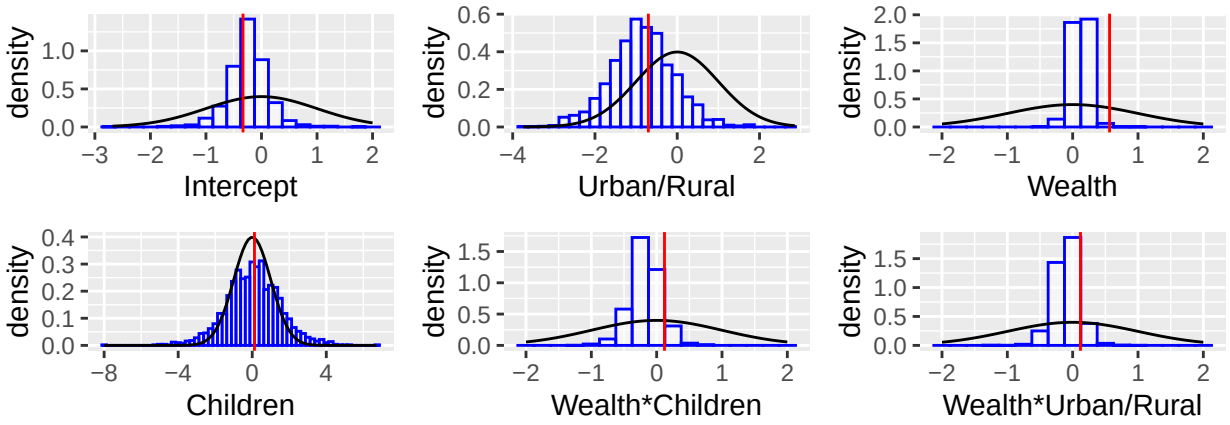

Figure 8: Prior and posterior comparison of the regression parameters. The prior density is given (black) along with posterior histogram (blue), and the likelihood-based estimate from lme4 (red).

Prediction of average woman who lives in urban area from an unknown care region.

The average values of independent variables from urban areas (the average urban woman) and used to give a probability of birth control uptake using the hierarchical model.

Table 3 gives the parameter estimates for this observation. The rhat and n_eff shows good convergence (rhat = 1 and n_eff over 100 represents good convergence).

Our regression value gives a probability of 49% (5% Confidence Interval: (0%, 100%)) that an average woman in an urban area from an unknown care region uses birth control. For reference, the average rate of birth control from the raw data is 39.25%, so this predicted observation has an increases chance of taking birth

control, highlighting the fact that urban areas could be a good area for the company to expand their market. However, the confidence interval suggests this improved probability could be down to chance alone.

Table 3: Parameter estimates of the average urban woman from unknown region

	mean	se_mean	sd	2.5%	50%	97.5%	n_eff	Rhat
yTilde	0.49	0.01	0.5	0	0	1	1129.37	1

The posterior prediction of the birth control used by an average woman who lives in an urban setting in region 1

Table 4 shows the parameter estimates for this, showing good n_eff and rhat values for convergence.

The mean in the table represents a probability of 37% (5% Confidence interval (9%, 73%)) of uptake of birth control from the hierarchical model, so this region has a slightly lower probability of uptake than from an unknown region in an urban setting, suggesting the company may not want to focus their sales on this region as this falls below the average intake. The wide confidence interval suggests this probability could be down to chance alone.

Table 4: Parameter estimates for average urban woman in region 1

	mean	se_mean	sd	2.5%	50%	97.5%	n_eff	Rhat
roneTilde	0.37	0.01	0.48	0	0	1	1878.97	1

The posterior prediction of the birth control use by an average woman who lives in an urban setting in the regions 14

Table 5 shows the parameter estimates for this, showing good n_eff and rhat values for convergence.

The mean in table 5 represents a probability of 92% (5% confidence interval (76%, 99%)) of uptake of birth control, so this region has a much higher probability of uptake than from an unknown region, and a much higher probability of uptake than region 1. The confidence interval suggests there is strong evidence that the probability of uptake for this profile of woman.

All this suggests the company should focus on selling in this region if it wants to expand it's market.

Table 5: Parameter estimates for average urban woman in region 14

	mean	se_mean	sd	2.5%	50%	97.5%	n_eff	Rhat
rfourtTilde	0.92	0.01	0.28	0	1	1	1823.77	1

Appendix

Hierarchical model diagnostic metrics for future model comparison

```
## Warning:
```

```
## 13 (0.7%) p_waic estimates greater than 0.4. We recommend trying loo instead.

##
## Computed from 2000 by 1934 log-likelihood matrix
##
##           Estimate   SE
## elpd_waic  -1244.0 18.5
## p_waic       114.0  4.0
## waic        2488.0 37.0
##
## 13 (0.7%) p_waic estimates greater than 0.4. We recommend trying loo instead.
```

Uninformative prior regression results

Citations

- Lambert, B. (2018). A Student's Guide to Bayesian Statistics (1st ed.). SAGE Publications Ltd.
- Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30.
- Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963
- Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595
- Vehtari A, Gabry J, Magnusson M, Yao Y, Bürkner P, Paananen T, Gelman A (2020). "loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models." R package version 2.4.1, <URL: https://mc-stan.org/loo/>.
- Simon Garnier (2018). viridis: Default Color Maps from 'matplotlib'. R package version 0.5.1. https://CRAN.R-project.org/package=viridis
- Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra
- Xavier Fernández i Marín (2016). ggmcmc: Analysis of MCMC Samples and Bayesian Inference. Journal of Statistical Software, 70(9), 1-20. doi:10.18637/jss.v070.i09
- Gabry J, Mahr T (2021). "bayesplot: Plotting for Bayesian Models." R package version 1.8.0, <URL: https://mc-stan.org/bayesplot/>.
- Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2. http://mc-stan.org/.
- Jonah Gabry (2018). shinystan: Interactive Visual and Numerical Diagnostics and Posterior Analysis for Bayesian Models. R package version 2.5.0. https://CRAN.R-project.org/package=shinystan
- Hadley Wickham (2020). modelr: Modelling Functions that Work with the Pipe. R package version 0.1.8. https://CRAN.R-project.org/package=modelr
- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.