Global Distribution of Unintended Pregnancies and Contraceptive Usage

Taylor Glass and Aryln Santiago

2024-04-17

# 1. Summary/Abstract

Our study investigates the association between unintended pregnancy rates and short-acting reversible contraception (SARC) use across low- and middle-income countries. Data from the Guttmacher Institute’s Adding It Up project was used, which included information on reproductive-aged women (15-49) and various demographic factors. Single and multivariate linear regression models were employed to examine relationships between predictor variables and the outcome variable (unintended pregnancy rate).

The exploratory data analysis revealed a less skewed distribution of unintended pregnancies among SARC users compared to other contraceptive methods. This led researchers to focus on SARC use for the remainder of the study. The Latin America/Caribbean region displayed the strongest association with unintended pregnancy rates, followed by Asia. The Oceania and Europe region had weaker associations. Marriage status, both currently and never married, was inversely associated with unintended pregnancy rates. This means that as the percent of women who are currently married or never married increases, the unintended pregnancy rate decreases. Maternal mortality rates showed a positive correlation with the outcome.

A full model incorporating five predictors (total SARC users, region, marital status, and maternal mortality rates due to unintended pregnancy and abortion) was built. Despite performing the best among the evaluated models and being validated through cross-validation and the bootstrap method, the full model exhibited tendencies towards overfitting the data.

Lasso regression, a technique that performs variable selection, was employed to address these overfitting concerns. The analysis identified the total number of SARC users as the strongest predictor, followed by region. Marital status and maternal mortality rates displayed weaker negative associations with unintended pregnancy rates. While the direction of the association between abortion-related maternal mortality and unintended pregnancy rates flipped in the Lasso model, the coefficient (effect size) remained minimal.

Overall, this study highlights the influence of regional variations, short acting reversible contraception use, and marital status on unintended pregnancy rates in low- and middle-income countries. The findings suggest a need for further investigation into the factors influencing maternal mortality rates and unintended pregnancies.

# 2. Introduction

## 2.1 General Background Information

We will be looking for patterns in contraceptive methods and higher unintended pregnancies across all age groups (15-19, 20-24, 25-35, and 35-49) to investigate the effectiveness of three contraceptive methods: long acting reversible and sterilization,short acting reversible, and traditional. Financial, social, religious, and cultural factors are current challenges against efforts towards reducing unintended pregnancies (Jonathan Marc Bearak et al., 2022). According to a cross-sectional survey done in 2017, contraceptive use was prevalent among women who had attained higher education, were in stable partnerships, and identified with the dominant ethnic group . Additionally, factors like previous pregnancies and immigration status did not appear to influence contraceptive use in this study (Rivet, Robinson, Sydora, & Ross, 2021). African and Latin American/Caribbean countries have the highest percentages of unintended pregnancies. This could be due to the current legal status of contraception within Latin American countries. Free emergency contraception access exists, but regulations differ. Chile, Colombia, and Ecuador legally recognize access, while Nicaragua and Bolivia rely on Ministerial protocols. Argentina and Brazil lack legal recognition, but offer protocols and guides. Mexico requires provision to victims of sexual and domestic violence (Hevia, 2012). Stable contraceptive use among high-risk women suggests reaching even lower national unintended pregnancy rates might be challenging (Kavanaugh & Jerman, 2018).

## 2.2 Description of data and data source

The data source for this project is the Guttmacher Institute, which is a well-known research organization focusing on improving sexual and reproductive health around the world. The data used here is from their Adding It Up project. The goals of this project include estimating the need for, impact of, and costs associated with providing sexual and reproductive health services. Low and middle income countries are the target audience for this research, specifically women of reproductive age (15-49). Some of the variables included in this dataset are modern contraceptive use, unintended pregnancies, unplanned births, and abortions averted. Observations are drawn from nationally representative surveys, including Demographic and Health Surveys, UNICCEF Multiple Indicator Cluster Surveys, US Centers for Disease Control Reproductive Health Surveys, Performance Monitoring for Action Surveys, and others. The UN Population Division is the source of population projections for women in this age group, and estimates of unintended pregnancies are from the Guttmacher Institute, WHO, and other authors, which have been adjusted to 2019.

For the purpose of this project, we chose several numeric variables from the data source: percentages of pregnancy outcomes, rates of safe and unsafe abortions, percentage of various contraceptive usages, percentages of care received for various pregnancy complications, and costs for many variables including abortions and STIs. There are only three character variables that can be used to group the numeric variables: country name, region, and sub region.

## 2.3 Questions/Hypotheses to be addressed

Is there a significant difference in the percentage of unintended pregnancies among short acting reversible contraception users among region?

The main outcome variables will be total percentage of unintended pregnancies among women using short acting reversible contraception methods (pct\_upreg\_sarc).

The predictor variables will include the total number of people women using each type of birth control standardized by population size, region, percentage of women currently married, percentage of women never married, total rate of maternal deaths, rate of maternal deaths due to unintended pregnancy and rate of maternal deaths due to abortion.

A full list of variables can be found in the processing-file in the code > processing-code folder.

# 3. Methods

Reproducing this project requires R, RStudio, and Microsoft word. The README files in each folder contain descriptions of the documents located in that folder to help guide users through the reproducible processes. Begin with the processing-file.qmd in the code > processing-code folder. This file loads the original dataset from the Guttmacher Institute’s Adding It Up project, loads the codebook for the dataset, and glimpses at the data. We chose 30 variables to clean to find the percentage of unintended pregnancies based on each type of contraceptive methods: long acting reversible and sterilization methods, short acting reversible methods, traditional methods, and no contraceptive usage. After the missing data was explored, we chose to use the percentage of unintended pregnancies variables related to each method of birth control because the total variable had unexplained missingness. The processed data is saved in the data > processed-data folder as processeddata.rds.

The next step is exploratory data analysis. The eda.qmd file in the code > eda-code folder contains all the code for this process. We produced a couple summary tables before we began a simple exploration of trends in percentage of unintended pregnancies by country in our first set of histograms. The second set of histograms explore contraceptive usage methods by country. These graphs revealed outliers, so we standardized by population size for each country. The next sets of plots are line graphs showing total usage of contraceptive methods to percentage of unintended pregnancies associated with that method. The last two sets of plots contain line graphs of the relationship between the two variables stratified by region and violin plots displaying the relationship between percentage of unintended pregnancies and the categorical usage variable created during processing. All of these figures are displayed and summarized in the Supplementary-Material.qmd file in the products > manuscript > supplement folder. This process led us to focus on the percentage of unintended pregnancies among short acting reversible methods as our single outcome because this method is becoming increasingly popular, so we feel that results from this contraception method will be the most relevant. Additionally, the exploratory data analysis process showed that the distribution of this outcome variable is less skewed than the other options, and the stratification by region is clearer, which allows for better understanding. All exploratory figures are saved in the results > figures > exploratory-figures folder.

Step three is the statistical analysis and model fitting portion found in the statistical-analysis.qmd file in the code > analysis-code folder. We explore the single outcome pct\_upreg\_sarc based on 7 predictors with simple linear regression, total number of women using short acting reversible methods, region, percent of currently married women, percent of women who never married, total rate of maternal mortality, rate of maternal mortality by unintended pregnancy, and rate of maternal mortality by abortion. We also explore the differences in the outcome between regions using an ANOVA test and Tukey’s Honestly Significant Difference test for pairwise comparisons within ANOVA. Result tables for each model and test are stored in the results > tables folder.

Step four is to critically analyze the performance of our models to determine which one performs the best in predicting our outcome, which is also done in the statistical-analysis.qmd file in the code > analysis-code folder. We chose 2 performance metrics to find for each of our 9 models: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). We chose these performance metrics because they are commonly used for linear regression models like the ones we have created. RMSE provides an average magnitude of the error in the predictors, and MAE provides an average absolute difference between the predicted and actual values. We include both metrics to increase the validity of our results. Once we decided on the final model, we use cross-validation to determine how well it will perform on data it has not seen before. This involves splitting the data into testing and training pieces , which allows us to compare the predictions created using both datasets. To examine the uncertainty in the model, we will use the bootstrapping method to resample our data 100 times to find confidence intervals for our predictions. This process is similar to cross-validation because it relies on resampling the training data. Lastly, we used a LASSO regression to add a penalty to each parameter included in the model to determine if any variables should be removed to improve model performance. It is important to tune the hyperparameters associated with LASSO regression to determine which model fits best. We use a grid search to ensure that we try all combinations of the parameters in our grid. We complete the tuning process with a 5-fold cross-validation repeated 5 times. We added a random forest model to attempt to improve model performance. We used another parameter tuning grid to allow for cross-validated tuning of 2 variables: mtry, the number of variable to possibly split at in each node, and min\_n, the minimal node size. We set the number of trees to 300 because it strikes a good balance between achieving the best performance metrics and the processing time required to do so. Each of these processes are crucial in determining which model performs best for our dataset. Result tables for each model and test are stored in the results > tables folder. Figures created from cross-validation, boostrapping, LASSO regression, and random forest modeling are stored in the results > figure > final-figures folder.

The Manuscript.qmd file for this document can be found in the products > manuscript folder.

## 3.1 Data aquisition

This research aims to analyze the factors influencing unintended pregnancies and contraception use in low-and-middle-income countries. The Guttmacher Institute’s Adding It Up project provides the foundation, offering data on sexual and reproductive health services in these regions.Focusing on women of reproductive age (15-49), the study usese information from nationally representative surveys and population projections to estimate unintended pregnancy rates.

While data on the overall percentage of unintended pregnancies was unavailable, we strategically utilized data on rates related to specific birth control methods. This approach allows us to analyze the relationship between these rates and various social and health factors. Factors like the number of women utilizing different birth control methods, maternal death rates, and marital status will be meticulously examined. All 7 of these predictor variables were found within the study’s codebook, which can can be found in the Data folder of this project. By discovering these relationships, the research aims to shed light on the complex landscape of contraception use and unintended pregnancy in low-and -middle-income countries. This knowledge can empower policymakers and healthcare providers to make informed decisions that ultimately improve reproductive health outcomes for women.

## 3.2 Data import and cleaning

As shown in the processing-file.qmd in the code > processing-code folder, we import the data from a .csv file that was downloaded directly from the Guttmacher AIU project website. We also included the codebook in this file. After taking a glimpse at the data, we identified the variables we need to clean and created a new dataset with those variables only. We explored missingness in the data, which is further explained in the supplementary-material.qmd file found in the products > manuscript > supplement folder. Simple processing tasks include converting variable types, converting decimal value into true percentages, creating variables to find total usage of each birth control method across age groups, exploring outliers, and creating new standardized total usage values based on population size.

## 3.3 Statistical analysis

We began with a simple linear regression model with each of the 7 predictor variables to determine the association betwen each variable and the outcome. We explored the region variable further because our main goal is to understand the difference in outcome by region. We used ANOVA to determine the difference between regions, and we added Tukey’s Honeslty Significant Difference test to examine pairwise associations. This led to discovering correlations between some of the predictors, which we examined through correlation coefficients. We tried out a combination of these correlated predictors in 2 different models to further examine the interactions. After considering which predictors are most impactful on our outcome, we narrowed the final predictors down to a list of 5. We included a multiple linear regression model with all 5 of these predictors. We used the RMSE and MAE performance metrics to evaluate all 9 models that we created to determine which one fit best. Lastly, we used cross-validation to determine how well our multiple linear regression model performs on data it has not seen before. We also used LASSO regression and random forest modeling as forms of subset selection by applying a regularization penalty and generating a tree structure, respectively. Lastly, we examine the AIC values associated with our two best performing models.

# 4. Results

## 4.1 Exploratory/Descriptive analysis

After deciding on a single outcome variable, we created a summary table with that outcome and the seven predictors we will use to predict it, which can be seen in [Table 1](#tbl-1).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Table 1: Summary Table Of Numeric Variables   |  | Region | | | | | | --- | --- | --- | --- | --- | --- | | **Predictor Variable** | **Africa**, N = 53 | **Asia**, N = 36 | **Europe**, N = 11 | **Latin America & the Caribbean**, N = 24 | **Oceania**, N = 8 | | pct\_upreg\_sarc |  |  |  |  |  | | Mean | 13 | 14 | 23 | 24 | 6 | | Range | 3, 36 | 3, 43 | 2, 39 | 9, 48 | 3, 12 | | pct\_currentlymarried |  |  |  |  |  | | Mean | 60 | 66 | 60 | 54 | 60 | | Range | 33, 85 | 44, 82 | 53, 73 | 32, 64 | 51, 66 | | pct\_nevermarried |  |  |  |  |  | | Mean | 31 | 29 | 29 | 35 | 34 | | Range | 12, 60 | 14, 50 | 16, 38 | 19, 66 | 26, 42 | | rate\_matdeaths |  |  |  |  |  | | Mean | 420 | 96 | 12 | 95 | 79 | | Range | 37, 1,155 | 7, 638 | 2, 19 | 25, 480 | 34, 145 | | rate\_matdeaths\_upreg |  |  |  |  |  | | Mean | 1,656 | 701 | 90 | 208 | 318 | | Range | 109, 7,952 | 35, 4,090 | 19, 203 | 48, 839 | 115, 490 | | rate\_matdeaths\_abs |  |  |  |  |  | | Mean | 160 | 17 | 9 | 23 | 40 | | Range | 7, 609 | 1, 48 | 3, 30 | 3, 74 | 13, 85 | | Number of women of reproductive age, 15-49 |  |  |  |  |  | | Mean | 5,956 | 30,567 | 5,232 | 6,832 | 350 | | Range | 50, 46,217 | 113, 353,357 | 146, 34,152 | 28, 57,089 | 26, 2,213 | |

We created a multitude of exploratory plots to understand our data better. The first exploratory result shown in [Figure 1](#fig-1) displays the distribution of the outcome variable: percent of unintended pregnancies due to short acting reversible methods. The pct\_upreg\_sarc variable is slightly skewed to the right with an average around 12%.

|  |
| --- |
| Figure 1: Number of Countries With Percent Unintended Pregnancy Using SARC |

The next set of variables we explored were the total counts of women using each birth control method. There were two large outliers that prevented the graphs from being useful. After exploring the outliers, we standardized using the population size of reproductive aged women (15-49 years). There is a positive association between the percentage of unintended pregnancies among short acting reversible method users and the total number of women using that method of birth control, which is to be expected. Our outcome variable is the number of unintended pregnancies among women who are using short-acting reversible contraception, so the number of unintended pregnancies that can be classified in this outcome is dependent on the number of women using this type of contraception. This percentage should increase as the number of women using short acting-reversible contraception increases because more women are at risk of experiencing an unintended pregnancy due to this contraception if they are using this type of contraception in the first place. This outcome variable does not speak to the effectiveness of short-acting reversible contraception, so it makes sense that our outcome would increase as more women put themselves at risk of experiencing the outcome through becoming classified as short-acting reversible contraception users. As explained in the supplementary material, the dataset includes the percentage of unintended pregnancies among long-acting reversible contraception and sterilization users, traditional birth control users, and users who do not use a method of birth control. Each of these percentages are positively associated with number of people in its respective classification group that it represents. The association for the contraception that we chose to focus on for this analysis, short acting reversible contraception, is displayed in [Figure 2](#fig-2).

|  |
| --- |
| Figure 2: Graph Depciting Percent Unintended Pregnancy by Standarized Count of Total Women |

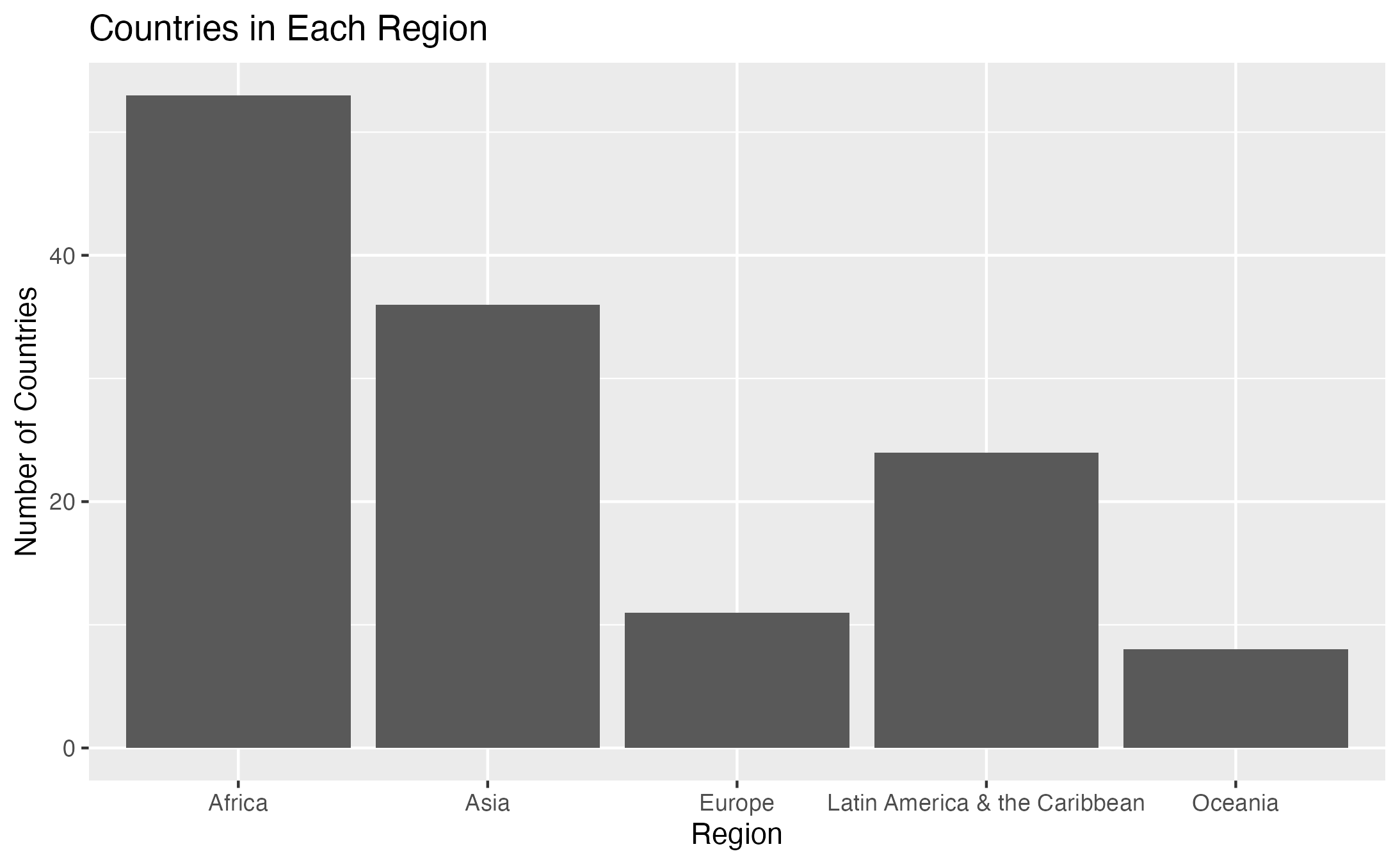
We stratified these four graphs by region to examine how the relationship differs across each of the regions which is shown in [Figure 3](#fig-3) . The graph for short acting reversible methods shows clearer differences between regions, and Latin America/the Caribbean leads in the percentage of unintended pregnancies in this category, but Africa shows the sharpest increase in percentage of unintended pregnancies as the total population increases.

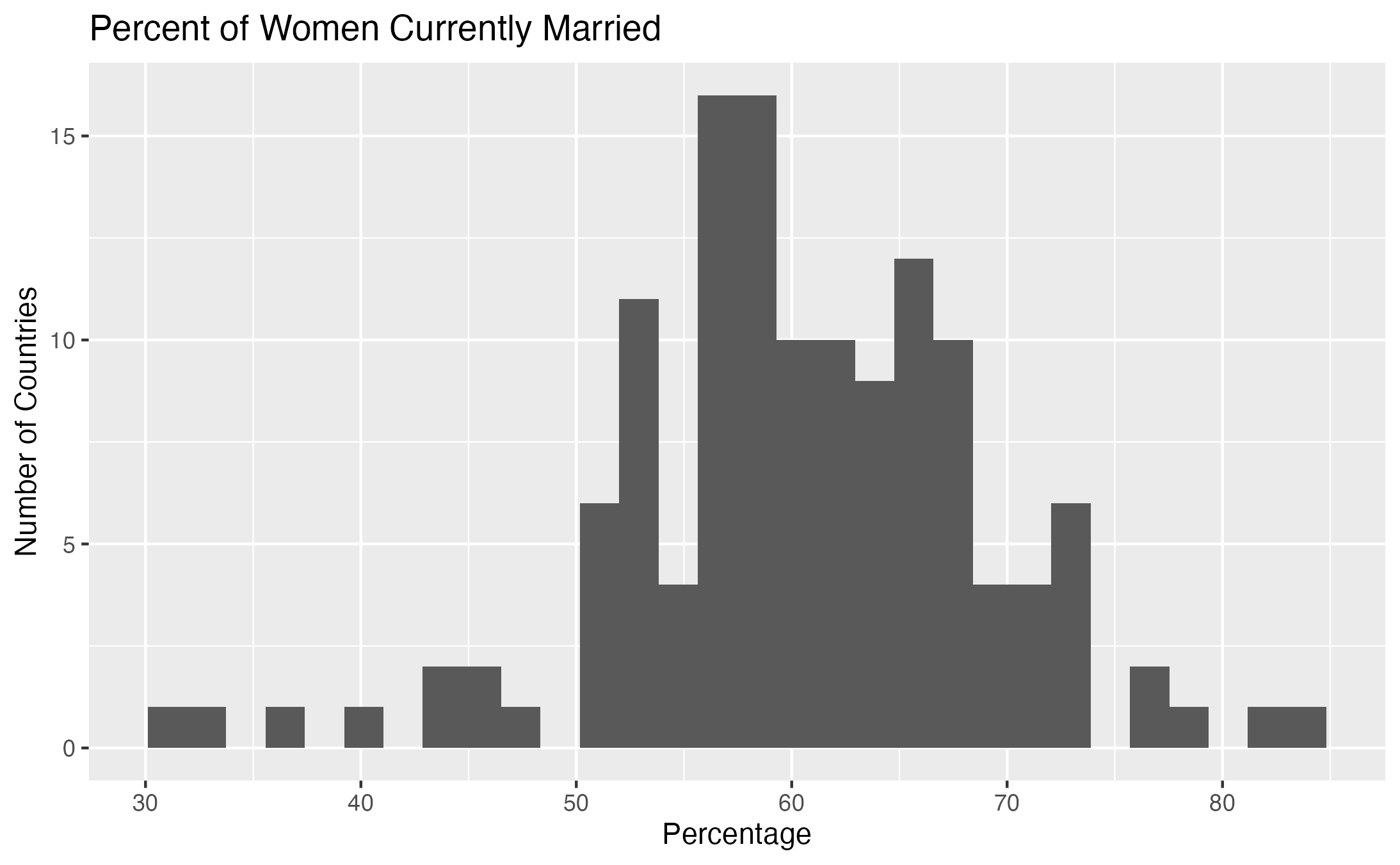
|  |
| --- |
| Figure 3: Graph Depicting Percent Unintended Pregnancies Using SARC by region |

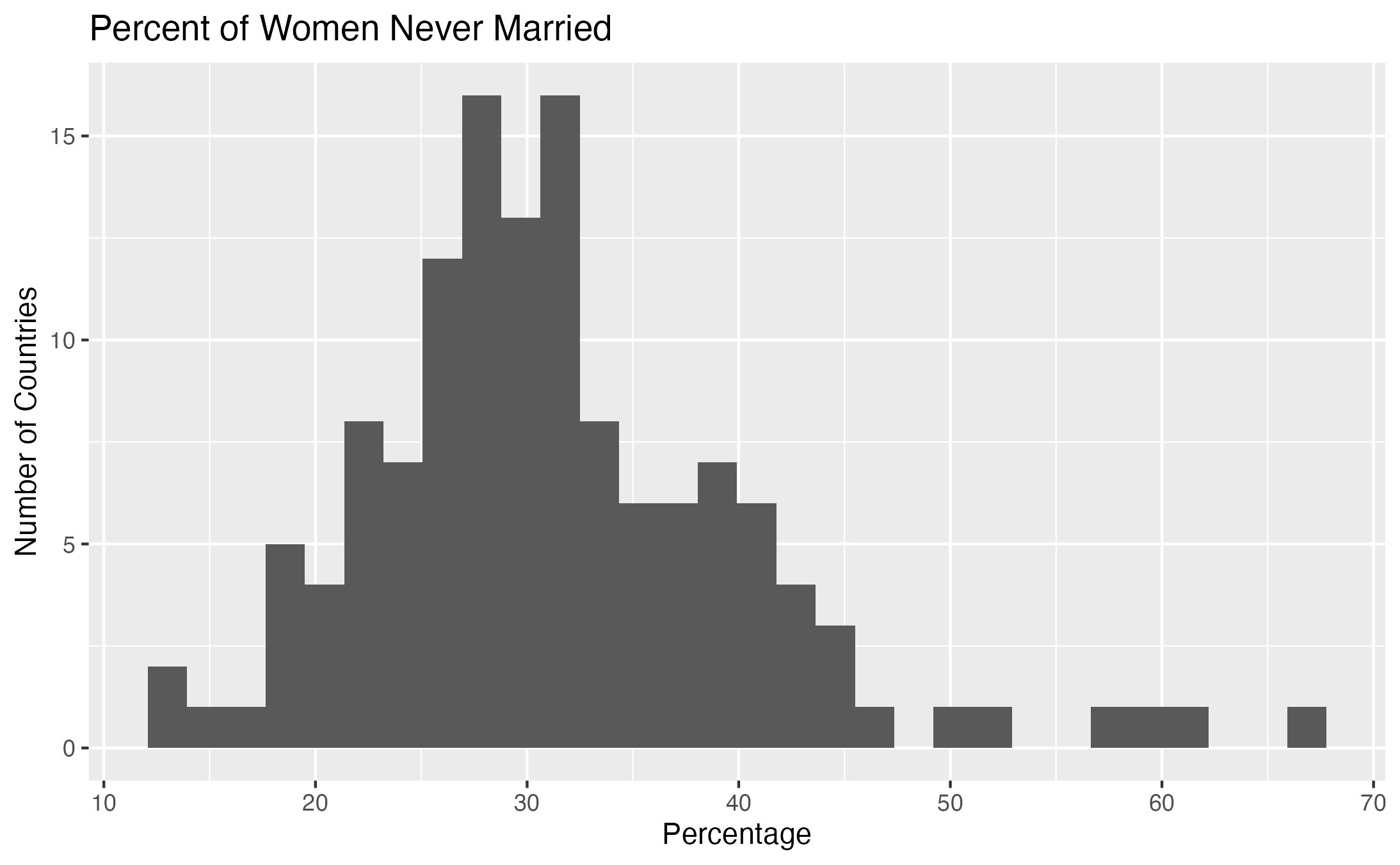
We also created violin plots using the categorical usage variables for each birth control method as shown in [Figure 4](#fig-4). It appears that the distribution of short acting reversible method usage is skewed towards lower percentages of unintended pregnancies across all usage levels, suggesting that more countries have lower percentage of unintended pregnancies among women using this method. The high usage countries appear to have a bottleneck in percentage of unintended pregnancies. The high and low usage countries appear to be skewed towards higher levels of unintended pregnancies, but the medium usage countries appear to have lower percentage of unintended pregnancies.

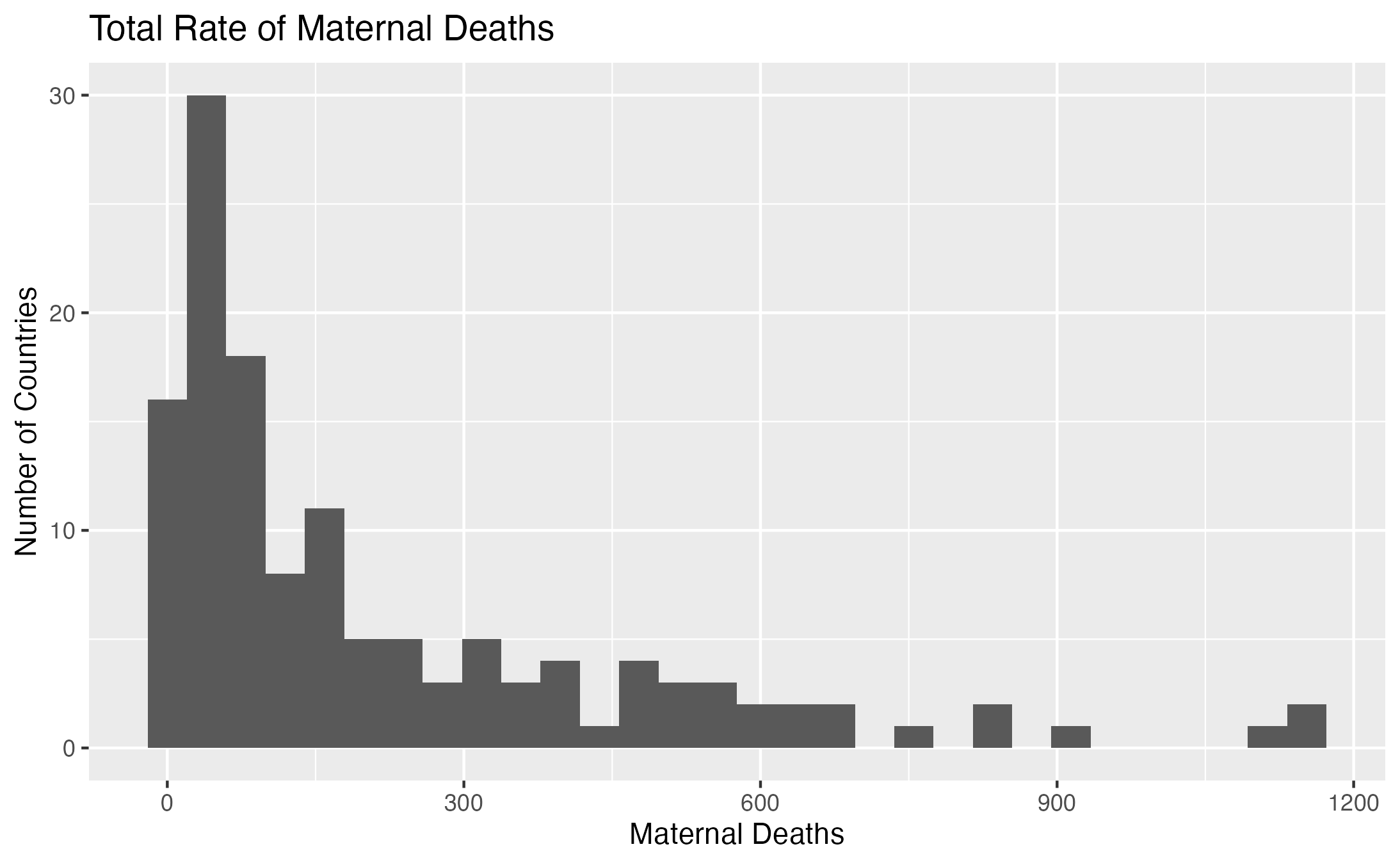
|  |
| --- |
| Figure 4: Violin Plot Depicitng Usage of SARc Per Region |

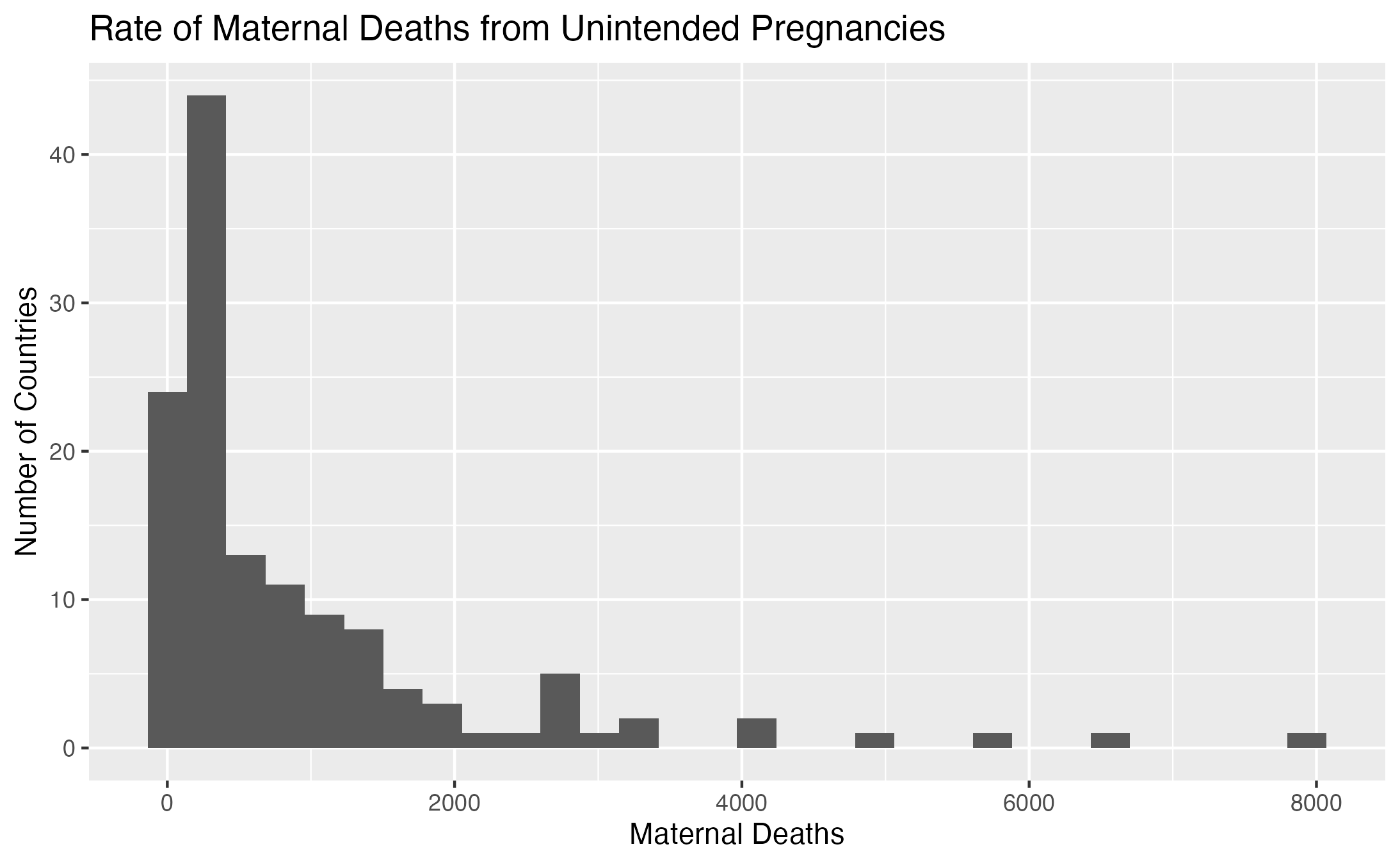
Lastly, we explored the predictor variables that will be used in our models, and the graphs are shown in **?@fig-5** - **?@fig-10**. The Africa region accounts for most of the observations with about 53 countries, followed by Asia that includes about 35 countries. Latin America/the Caribbean accounts for about 25 countries. Europe and Oceania account for the least number of observations. The distribution of women currently married is approximately normal with an average around 57% and no major outliers. The distribution of women never married is slightly skewed to the right with an average around 30%, but there are a few outliers greater than 60%. The distribution of the total maternal mortality rate is highly skewed to the right with a couple outliers. We would have expected the highest rates to belong to India and China because they have the most people, but it is Chad, Sierra Leone, and South Sudan that account for the highest rates of total maternal mortality, which each belong to the African region. The distribution of the maternal mortality rate from unintended pregnancies is highly skewed to the right with a couple outliers, which is to be expected based on the trend shown in the previous histogram. Chad and South Sudan also account for the highest rates of maternal mortality in this category. The distribution of maternal mortality rate from abortions is also highly skewed to the right. The outliers are accounted for by Gambia and Mauritania in this predictor, which both belong to the African region.

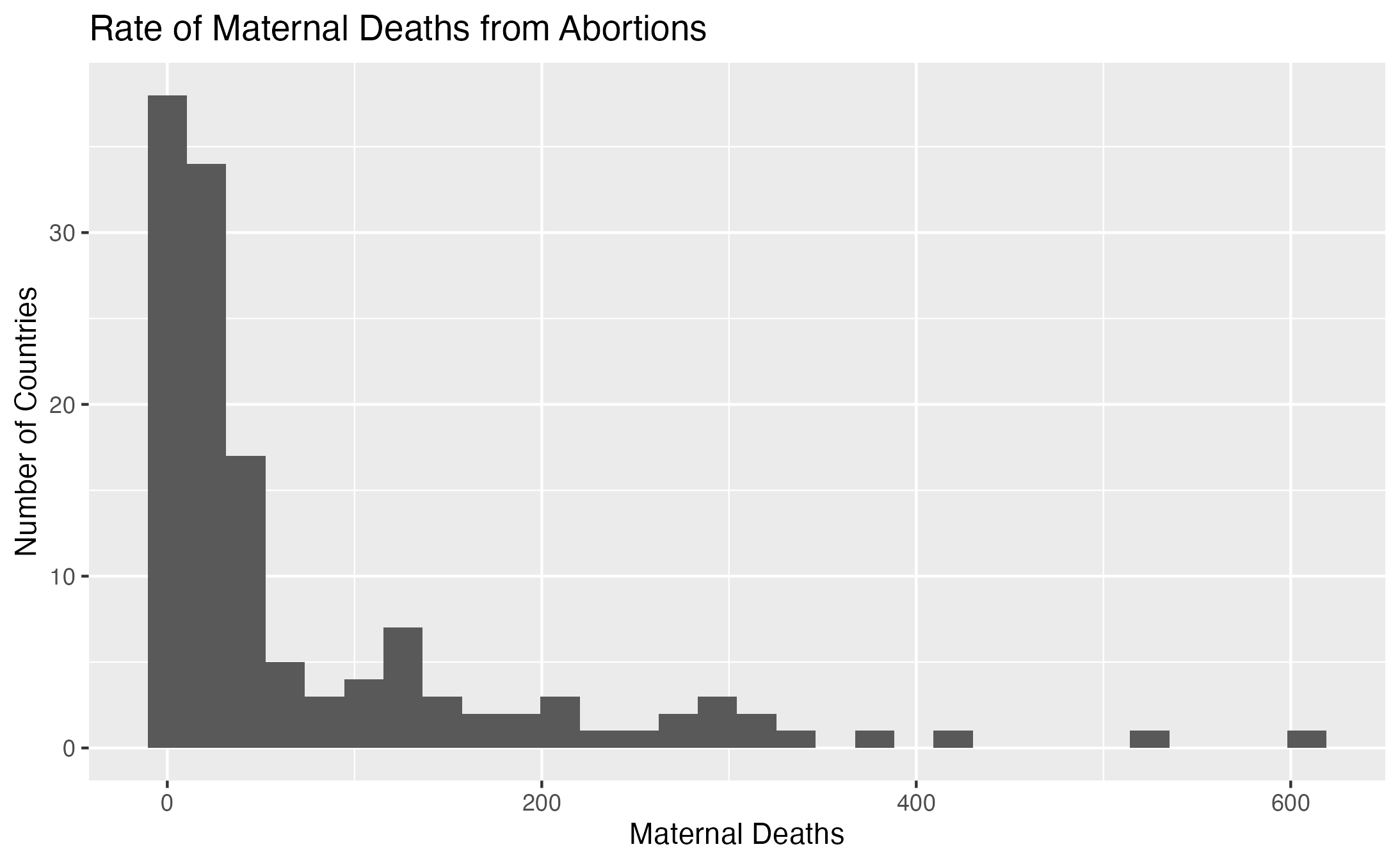
{#fig-5 - fig 10-1}

{#fig-5 - fig 10-2}

{#fig-5 - fig 10-3}

{#fig-5 - fig 10-4}

{#fig-5 - fig 10-5}

{#fig-5 - fig 10-6}

## 4.2 Basic statistical analysis

We completed a simple linear regression analysis with each of the 7 predictors shown in [Table 2](#tbl-2). The predictor with the strongest positive relationship with the outcome is total number of women using short acting reversible methods (sarc\_standard). This percentage should increase as the number of women using short acting-reversible contraception increases because more women are at risk of experiencing an unintended pregnancy due to this contraception if they are using this type of contraception in the first place.

|  |
| --- |
| Table 2  # A tibble: 2 × 5  term estimate std.error statistic p.value  <chr> <dbl> <dbl> <dbl> <dbl> 1 (Intercept) -1.70 0.947 -1.80 7.42e- 2 2 sarc\_standard 87.7 4.16 21.1 1.01e-43 |

The second model we created used the region predictor, which revealed that the Latin America/the Caribbean region had the strongest relationship with the outcome variable followed closely by the Asia region. The Oceania region does not appear to have a strong effect on the outcome variable with the Europe region only showing a slightly larger association. These associations can be seen in [Table 3](#tbl-3).

|  |
| --- |
| Table 3  # A tibble: 5 × 5  term estimate std.error statistic p.value  <chr> <dbl> <dbl> <dbl> <dbl> 1 (Intercept) 0.418 0.159 2.63 0.00960  2 regionAsia 1.47 0.250 5.87 0.0000000350 3 regionEurope 0.330 0.384 0.860 0.392  4 regionLatin America & the Caribbean 1.62 0.285 5.69 0.0000000827 5 regionOceania 0.251 0.439 0.572 0.568 |

Details about the 10 different models we created can be found in the statistical-analysis.qmd file in the code > analysis-code folder along with an exploration of the correlation between predictors.

## 4.3 Full analysis

We created one model with all 5 predictors using multiple linear regression, shown in [Table 4](#tbl-4). This model shows the strongest predictor is the number of women using short acting reversible contraception. The next strongest predictor is the European region followed by the Latin America/the Caribbean region and the Asia region, which is a shift from the model with the singular region predictor. The Oceania region now has a negative association with the outcome. The currently married predictor is associated with a weak decrease in the outcome. The coefficient estimates for the rate of maternal death due to unintended pregnancies and abortion both show a very weak positive association with the outcome.

|  |
| --- |
| Table 4  # A tibble: 9 × 5  term estimate std.error statistic p.value  <chr> <dbl> <dbl> <dbl> <dbl> 1 (Intercept) -2.66 3.60 -0.740 4.60e- 1 2 sarc\_standard 83.5 4.34 19.2 7.58e-39 3 regionAsia 4.80 1.30 3.69 3.30e- 4 4 regionEurope 6.94 1.68 4.12 6.77e- 5 5 regionLatin America & the Caribbean 5.29 1.29 4.10 7.51e- 5 6 regionOceania -0.0328 1.91 -0.0172 9.86e- 1 7 pct\_currentlymarried -0.0156 0.0554 -0.282 7.79e- 1 8 rate\_matdeaths\_upreg -0.000130 0.000428 -0.304 7.62e- 1 9 rate\_matdeaths\_abs 0.000180 0.00542 0.0332 9.74e- 1 |

To determine which model performed the best, we started by examining the RMSE and MAE performance metrics. We found that the model with the standardized total count of women using the short acting reversible contraception method performed the best in a single predictor model compared to the percent of currently married women and the rate of maternal mortality due to abortion. [Table 5](#tbl-5) shows the RMSE and MAE values for each of these three models compared to the null model with no predictors. The model with the population size using SARCs is the only single predictor model that performs better than the null. The full model with all 5 predictors performs the best with the lowest RMSE and MAE values.

|  |
| --- |
| Table 5  RMSE MAE Population Size Using SARCs 5.143319 3.953767 Percent Currently Married 10.175909 8.107884 Rate of Maternal Deaths due to Abortion 10.141275 8.216870 Full Model with 5 Predictors 4.361363 3.386193 Null Model 10.801787 8.216870 |

We used cross-validation of the two best models to determine how well the model performs on data it has not seen yet by splitting the data into testing and training pieces. The cross-validated results, shown in [Table 6](#tbl-6) lead to the same conclusion: the multiple linear regression model with 5 predictors performs better. The RMSE value for the single predictor model with sarc\_standard is 4.822226, which is lower the the first RMSE value of 5.143319. While the RMSE value for the multiple linear regression model is slightly higher than the first time it was calculated, it is still lower than the single predictor model. Additionally, the standard error of the RMSE for the multiple linear regression model is 0.2768, which is lower than the standard error of the RMSE for the single predictor model of 0.3528. This means there is higher variance in the single predictor model. The R-squared value of 0.8206 for the multiple linear regression model is higher than 0.7861 for the single predictor model, which means the multiple linear regression model accounts for more of the variability in the data. Overall, the multiple linear regression model performs better.

|  |
| --- |
| Table 6  RMSE Standard Error R-Squared Standard Error Single Predictor Model 4.82222600 0.25279125 0.78606700 0.04292202 Full Predictor Model 4.40571600 0.27679502 0.82060000 0.02815933 |

Since we determined that the full model with all 5 predictors performs best, we graphed the predicted versus observed values using this model with both the testing and training data to further examine model performance. As seen in [Figure 5](#fig-11), the training model predictions, shown in teal, hover pretty closely around the expected relationship between the outcome and the 5 predictors, which is depicted as the 45 degree line. The testing model predictions, shown in coral, are much more extreme than the training model predictions because they are farther from the 45 degree line. While there is some overlap between the two sets of predictions, it is likely that our model is overfitting the data.

|  |
| --- |
| Figure 5: Evaluating the final model performance on test data reveals that the model is strongly overfitting the data. The testing model predictions are shown in coral, and the variation in these observations is much higher compared to the training model predictions, shown in blue. Ideally, there would be complete overlap between the testing and training data predictions following the black dotted line that represent the ideal relationship between observed and predicted values. The majority of testing data predictions do not overlap with the training data prediction in this case, and the testing data predictions are father away from the preferred relationship shown by the black dotted line. |

We went one step further to examine the uncertainty in the full model predictions using the bootstrap method with 100 resamples of the training data. The figure shown in [Figure 6](#fig-12) displays the 95% confidence intervals, shown as dark blue lines, for the predictions made by the model. The confidence intervals are narrow, which shows there is minimal uncertainty in the model prediction. For the predictions that do not fall on the expected 45 degree line, the confidence intervals do not intersect with the line either. These findings increase the validity of the conclusion that the model is overfitting the data because there is not much uncertainty in the predictions.

|  |
| --- |
| Figure 6: Evaluation of uncertainty in the full model predictions reveals decent overlap between the point estimates, shown in black triangles, and the median prediction observation, shown in coral circles. The error bars, shown in blue, represent 95% confidence intervals of the median estimates, and most of the confidence intervals contain the point estimates. This provides confidence in the certainty of the model’s predictions, but this is not very meaninful considering how the full model overfits the data. |

Since we have evidence that our model overfits the data, the last technique we want to use to explore our full model is Lasso regression. This technique could reveal predictors that should not be included in our final model, which could reduce the risk of overfitting. As seen in [Table 7](#tbl-7), the LASSO regression coefficients are much smaller than the first full model we created without the regularization penalty, but all 5 predictors remain in the model. The sarc\_standard predictor still has the strongest relationship with the outcome. The percent of women currently married, rate of maternal deaths due to unintended pregnancies, and rate of maternal deaths to due to abortion all have weak negative associations with the outcome. This process confirms that our final model should contain all 5 predictors.

|  |
| --- |
| Table 7  # A tibble: 9 × 3  term estimate penalty  <chr> <dbl> <dbl> 1 (Intercept) -3.13 0.1 2 sarc\_standard 82.9 0.1 3 regionAsia 4.19 0.1 4 regionEurope 6.28 0.1 5 regionLatin America & the Caribbean 4.87 0.1 6 regionOceania -0.00949 0.1 7 pct\_currentlymarried -0.0000458 0.1 8 rate\_matdeaths\_upreg -0.000137 0.1 9 rate\_matdeaths\_abs -0.000597 0.1 |

It is important to tune the hyperparameters associated with LASSO regression to determine the model that fits best, which is shown in [Figure 7](#fig-13). As the regularization penalty increases past 1, we see that the RMSE value increases exponentially. The lowest RMSE for the model is slightly greater than 5 when the regularization penalty is less than 1. When the regularization penalty increases to 10, the RMSE increases to almost 11. Since we originally chose 0.1 as our penalty parameter, we found one of the best performing models with the LASSO regression based on an RMSE value slightly greater than 5, which i comparable to the RMSE found in the first regression model with all 5 predictors.

|  |
| --- |
| Figure 7: Parameter tuning for the LASSO regression reveals that the parameter should be smaller than 1 to ensure the lowest RMSE performance metric value. The RMSE value increases exponentially when the penalty parameter exceeds 1, so this provides confidence in our choice of using 0.1 as the penalty parameter in the prevoiusly displayed LASSO regression model. There is no evidence to use a smaller penalty parameter than 0.1 because it does not result in a lower performance metric. |

After creating the random forest model, we found that the RMSE and MAE values are exactly the same because the same 5 predictors are used for the outcome, but the R-squared value shows that the random forest model performs slightly better. Since a higher R-squared value represents smaller differences between the observed and predicted values, we can conclude that this model is predicting our outcome with slightly higher accuracy. The plot of the predicted versus observed values shown in [Figure 8](#fig-14) displays stronger correlation for predicted values compared to the testing and training done with cross-validation previously.

|  |
| --- |
| Figure 8: The graph of predicted versus observed values in the random forest model shows less deviation from the expected relationship visualized by the black dotted line in comparison to the multiple linear regression model displayed previously. While there are still several outliers as the predicted values increase past a value of 20, the observations hover around the expected relationship with conistency for values less than 10. While it is still likely that the random forest model is overfitting because it uses the same predictors, it appears that this model has higher predictive accuracy. |

Again, it is crucial to tune hyperparameters associate with machine learning models. Based on the red line displayed in [Figure 9](#fig-15) that shows the lowest possible RMSE score, our random forest model performs the best with a node size of 1 and 4 predictors. This begs the question of which one of the 5 predictors included in our model should be dropped to improve model performance because all previous performance metrics suggested that including all 5 predictors created the best model. However, there is a minimal increase in RMSE when all five predictors are included in the model, but the results suggest that a full model should have a node size of 4 to achieve the lowest RMSE.

|  |
| --- |
| Figure 9: After tuning two parameters for the random forest model, it is clear that the best model based on the lowest RMSE performance metric value has a minimal node size of 1 for 4 out of the 5 models. This graph suggests that the final model should contain 4 predictors to achieve the best performance metric, but it does not provide information on which parameter should be dropped from the model. Interestingly, the best model with all 5 predictors has a minimal node size of 4. |

As seen in the statistical analysis file, we created 10 different model options with unique combinations of our 7 predictors. After considering effect sizes and performance metrics for each model, we have decided that the LASSO regression model with 5 predictors is the best overall choice for predicting our outcome of percentage of unintended pregnancies due to short acting reversible methods. After considering each of the 10 models we created throughout the statistical analysis process, the random forest model with 5 predictors is the best overall choice for predicting our outcome of percentage of unintended pregnancies due to short acting reversible methods. We used cross-validation with LASSO regression and random forest modeling to explore subset selection of our full model with all 5 predictors. The LASSO regression kept all 5 predictors, and the coefficients for the Asia, Europe, and Latin America/the Caribbean predictors were the largest behind the sarc\_standard predictor. The smallest coefficient was assigned to the percent of currently married women predictor. We suggest using the full model with all 5 predictors with a LASSO regression because it accurately portrays the importance of each predictor while providing a penalty to predictors that do not meaningfully contribute to the model. While this model has a slightly lower R-squared value compared to the random forest model, the evidence of overfitting is so strong that this difference is negligible. Cross-validation of the model with all 5 predictors showed that predictions from testing data deviated from the expected relationship with larger variance than the predictions from the training data. Since the model did not perform well when used on test data, we conclude that there is overfitting in our model. Unfortunately, it is likely that any of these models are not optimal choices for predicting our outcome based on the poor performance of our model on testing data. While we suggest using the LASSO regression with all 5 predictors to predict our outcome because it is slightly easier to interpret, future researchers who are willing to sacrifice interpretation for predictive power should use the random forest model with 4 predictors based on the slightly higher predictive performance shown by the R-squared performance metric.

# 5. Discussion

## 5.1 Summary and Interpretation

Our exploratory data analysis allowed us to examine the distribution of our outcome of interest, percentage of unintended pregnancies due to short acting reversible contraception, and the predictors used to describe the outcome. Most low- and middle-income countries were from the African region, followed by Asia and Latin America/the Caribbean. The Latin America/the Caribbean region experienced the highest percentage of unintended pregnancies followed by Europe and Asia. This association is likely explained by the fact that women in these regions have better access to short acting reversible contraception methods because we would have expected the African region to experience the highest rate of our outcome based on the number of countries included from that region. After standardizing the number of women who use short acting reversible contraception based on total population size, it is clear the women in the Latin America/the Caribbean region use this contraception method at the highest rates, followed my the Asian region. Further analysis of the region predictor in a linear model with the outcome confirmed this finding by showing that Latin America/the Caribbean has the strongest association with the outcome, followed by Asia. Asia accounted for the highest percentage of currently married women. The highest rate of maternal mortality due to abortion and highest rate of maternal mortality due to unintended pregnancies are both accounted for by the African region. This cannot be explained by having the highest number of reproductive age women because the Asian region has 6 times the number of reproductive age women compared to the African region. This discrepancy is likely due to poorer access to maternal healthcare in Africa compared to Asia.

Our full statistical analysis of the final model with 5 predictors included the standardized number of women using short acting reversible methods, region, percent of currently married women, rate of maternal mortality due to abortion, and rate of maternal mortality due to unintended pregnancies as predictors. Cross-validation using LASSO regression and random forest modeling techniques did not result in an increase in performance metrics, and the model is strongly overfitting the data based on the predicted versus observed outcomes between the testing and training data. Based on the model using LASSO regression, the number of women using short acting reversible contraception is the strongest predictor of the outcome, which is to be expected because women must be using this form of contraception to have an unintended pregnancy due to failure of this contraceptive method in the first place. The most influential factor in this model is the region variable with European, Asian, and Latin American regions having the strongest effects on the outcome. This finding implies that these three regions have the best access to SARCs because they also have the highest standardized population size using this form of contraception in the first place. Since the African region was always selected as the control for the model, it is likely that this region has the lowest access to SARCs. The percent of currently married women, rate of maternal mortality due to abortion, and rate of maternal mortality due to unintended pregnancies as predictors are all negatively associated with the outcome in this model. This implies that the percentage of unintended pregnancies decreases when a higher percentage of women are currently married and when the country experiences a higher maternal mortality rate based on abortion and unintended pregnancies. However, these effects are so small that it is likely these predictors do not have a real effect on the outcome. This analysis was successful in determining that region does have a significant effect on the percentage of unintended pregnancies due to short acting reversible methods with the Latin America/the Caribbean region experiencing the outcome at the highest rates.

## 5.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 5.3 Conclusions

*What are the main take-home messages?*

# 6. References

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

(Hevia, 2012), (Jonathan Marc Bearak et al., 2022), (Kavanaugh & Jerman, 2018), (Rivet et al., 2021)

Hevia, M. (2012). The legal status of emergency contraception in Latin America. *International Journal of Gynecology & Obstetrics*, *116*(1), 87–90. <https://doi.org/10.1016/j.ijgo.2011.10.008>

Jonathan Marc Bearak, Anna Popinchalk, Cynthia Beavin, Bela Ganatra, Ann-Beth Moller, Özge Tunçalp, & Leontine Alkema. (2022). Country-specific estimates of unintended pregnancy and abortion incidence: A global comparative analysis of levels in 2015–2019. *BMJ Global Health*, *7*(3), e007151. <https://doi.org/10.1136/bmjgh-2021-007151>

Kavanaugh, M. L., & Jerman, J. (2018). Contraceptive method use in the United States: Trends and characteristics between 2008, 2012 and 2014. *Contraception*, *97*(1), 14–21. <https://doi.org/10.1016/j.contraception.2017.10.003>

Rivet, D., Robinson, H., Sydora, B. C., & Ross, S. (2021). Contraception Use and Attitudes Among Women Seeking Pregnancy Termination: Descriptive Cross-Sectional Survey. *Journal of Obstetrics and Gynaecology Canada*, *43*(2), 197–203.e2. <https://doi.org/10.1016/j.jogc.2020.06.023>