Estimation of population at high risk of statelessness given data on the citizenship-obtaining procedure

Contents

1 Overview

The method described in this section produces an estimate of the size of the population who are at very high risk of statelessness by modeling the probability of success in the citizenship-obtaining procedure. The probability of success is informed by the characteristics of individuals who have undergone the procedure. Resulting estimated probabilities are re-weighted to give population-level estimates.

1.1 Case study

The case study used for this example is the Cote d'Ivoire, which has detailed survey data on the citizenship-obtaining procedure. In the Cote d'Ivoire report on statelessness, individuals are classified into four risk categories (no risk, low risk, high risk, and very high risk) based on their responses to the survey. Individuals are only placed in the high risk category if they have underwent the procedure to obtain citizenship and failed.

2 Data requirements

- a sample of individuals with information that can inform a risk level for statelessness, with characteristics (covariates) which may be relevant for predicting the risk level
- a sample of individuals who have undergone a citizenship-obtaining procedure, with characteristics which may be relevant for predicting their success in the procedure
- population counts for each possible combination of covariates, ideally from a census

2.1 The data in Côte d'Ivoire

In the survey conducted in Côte d'Ivoire, respondents are asked separately Q1. whether they have began a procedure to obtain a nationality (and the outcome, if applicable) Q2. whether they have applied for Ivorian citizenship via declaration between Jan. 2014 to Jan. 2016 (and the outcome, if applicable) Q3. whether they have applied for naturalisation (and the outcome, if applicable)

Respondents who answer "yes" to Q2 and Q3 are a subset of the respondents who answer "yes" to Q1. However, even if a successful outcome is reported for Q2 and Q3, they may still report an unsuccessful outcome for Q1. In order to reduce these responses into a single binary outcome, we make the following simplifications:

- Respondents are included in the data if they respond "Yes" to Q1 AND they have completed the procedure (the procedure is not recorded as "in progress") or they have a successful outcome in Q2 or Q3

- If a respondent reports a successful outcome for any of the three questions, their outcome is labelled as a success
- Otherwise, a respondent's outcome is labelled as a failure. This includes individuals who answer "No" to Q2 and Q3.

For example, respondents with the following responses would be labelled as successful outcomes because they report at least one method in which they were successful:

Q1: Yes, unsuccessful Q2: No Q3: Yes, successful

Q1: Yes, successful Q2: Yes, unsuccessful Q3: No

Q1: Yes, in progress Q2: Yes, successful Q3: Yes, unsuccessful

The following respondents would be labelled as unsuccessful outcomes:

Q1: Yes, unsuccessful Q2: No Q3: Yes, unsuccessful

Q1: Yes, unsuccessful Q2: No Q3: No

The following respondent would not be included in the outcome data:

Q1: Yes, in progress Q2: Yes, unsuccessful Q3: Yes, unsuccessful

3 Method

There are two main components to the model. In the first component, we model the risk levels of the individuals, and in the second component we model the probability of success.

3.1 High-level overview

1. An individual's risk level of statelessness (none, low, high) is modeled as a function of their birthplace and region
2. An individual's probability of failing the citizenship proceedure is modeled as a function of their risk level, birthplace and region
3. These probabilities are combined to give an overall risk of failure probability
4. The risk of failure probability is post-stratified using population counts by birthplace and region to obtain estimates of the count of persons at high risk of statelessness

3.2 Model for risk level

We model the risk level as a function of the geographic area and birthplace using a multinomial-logit regression setup. This is an extension of the usual binary logistic regression model, with three categories in this case (no risk, low, high). The outcome of this step is a set of predicted probabilities of individual's risk level of statelessness based on the country they were born in and the region where they live.

3.3 Model for citizenship procedure success

The next step is to model the probability of citizenship procedure success as a function of risk level, geographic region, and birth country. This is done using a logistic regression, with a binary yes/no outcome (failed citizenship procedure yes/no). The geographic region effect is modeled as a "random effect", which allows for prediction of the probability of citizenship procedure success to be made even for geographic regions that were not observed in the survey.

3.4 Producing population level estimates

Taken together, the previous two steps of the method allow for the probability of failing the citizenship procedure to be estimated for individuals based on birth country and geographic region. These probabilities are the combined with representative population-level counts by birth country and region (e.g. from a census) to obtain representative counts of those at high-risk of statelessness.

3.5 Assumptions

- individuals who underwent the citizenship obtaining procedure are comparable to those who have not begun a procedure
- individuals who are assigned "no risk" are indeed not at risk of statelessness

3.6 Possible extensions/improvements

- using census counts that are representative with respect to birthplace (the survey is only representative with respect to geographic area)
- additional covariates included in models
- this method only produces estimates for persons at high-risk of statelessness. Ideally this method would be combined with additional information about the link between failing citizenship procedures and true statelessness.

4 Step-by-step example

The following section walks through the proposed method using survey data from Cote d'Ivoire.

4.1 Empirical analysis

All calculations were down in the statistical software package R. The procedure requires packages `tidyverse`, `here`, `cmdstanr`, and `posterior`. The analysis involves a knowledge of statistics, multinomial and binomial logistic regression combined with post-stratification. All steps of the method are estimated using the probabilistic language Stan.

4.2 Script and model code

The data extraction and analysis relies on code which is in the project repo in the `code` folder. A summary of each is as follows:

- The `civ_prep.R` file processes the raw survey data. This takes as an input the raw survey data and saves a processed file called `survey2019_processed.rds`.
- The `civ.R` file contains code to load in and process data and run the models. This takes as an input the processed survey data from above, runs the model, and saves intermediate results as `civ.rds`.
- The `civ.stan` file is the Stan model code. This is required in the `civ.R` file.

4.3 Step 1: Read in survey data and extract required information

The `civ_prep.R` file takes the raw CIV survey data, extracts the relevant variables for analysis, recodes and renames if applicable, and saves the resulting file as an RDS file called `survey2019_processed.rds`.

4.4 Step 2: Prepare data for model

The next step is to get the processed survey data into a form that can be input into the statistical model. The code to do this is contained in the `civ.R` script. The first part of the code maps region, birth country and risk levels onto numeric values for the purposes of modeling. The data are then converted into matrices and a list in order to be inputted into the Stan statistical model.

4.5 Step 3: Run model to estimate probabilities by risk group, probability of citizenship procedure success, and get population counts, and save results

Line 140 of the `civ.R` file runs the Stan model. Once the model is run, summary statistics are calculated:

```r
q_probs <- c(0.025, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, 0.975)

# Calculate summary statistics for parameters
summary_table <- fit$summary(
  variables = NULL,
  mean, sd, rhat, ~quantile2(.x, probs = q_probs), ess_bulk, ess_tail
)
```

The inputs and outputs of the model are then saved to an intermediate file, which can be read in and manipulated in the next step:

```r
to_save <- list(
  time_to_fit = time_to_fit,
  summary_table = summary_table,
  data = data_list,
  region_dict = region_dict,
  birthplace_dict = birthplace_dict,
  pred_df = pred_df,
  draws_location = fit$output_files(),
  diagnostic_summary = fit$diagnostic_summary()
)

write_rds(to_save, here("output/intermediate/civ.rds"))
```

4.6 Step 4: Extracting results

We can now read in the intermediate results and assign it to an object called `civ_model_results`. This contains many different types of information, including model inputs and outputs. We can use the `names` function to see all the types of information within this object.

To start with, we extract the `pred_df`, which give populations (in the `n` column) by birth country and region of residence. This can be extracted from the full `civ_model_results` object using the `$` syntax.

```r
civ_model_results <- read_rds(here("output/intermediate/civ.rds"))
names(civ_model_results)
```

```
## [1] "time_to_fit"        "summary_table"      "data"
## [4] "region_dict"        "birthplace_dict"    "pred_df"
## [7] "draws_location"     "diagnostic_summary"
```

```
pred_df <- civ_model_results$pred_df
```

The model format allows results to be extracted in various forms. The summaries of results are given in the `summary_table`. Firstly, the total estimated count is called `total_count_pred`. We can extract this by searching for this name in the `variable` column of the summary table. The results show the mean estimate, standard deviation, and other information.

```
# Total stateless
civ_model_results$summary_table %>%
  filter(str_detect(variable, "total_count_pred"))
```

```
## # A tibble: 1 x 15
##   variable    mean     sd  rhat   q2.5     q5    q10    q25    q50    q75    q90
##   <chr>      <dbl>  <dbl> <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1 total_co~ 2.14e6 1.67e5  1.00 1.82e6 1.87e6 1.93e6 2.03e6 2.15e6 2.26e6 2.35e6
## # ... with 4 more variables: q95 <dbl>, q97.5 <dbl>, ess_bulk <dbl>,
## #   ess_tail <dbl>
```

To get populations and resulting counts by risk level, we can extract information on `pop_by_risk_level` and `counts_by_risk_level`. The former gives the estimated populations in the risk groups high (1), low (2), and none (3), and the latter gies the final counts at high-risk of statelessness. Note that the probability of statelessness for those who are estimated to have no risk is by assumption 0, so these counts go to 0.

```
# Estimated population by risk level
civ_model_results$summary_table %>%
  filter(str_detect(variable, "pop_by_risk_level"))
```

```
## # A tibble: 3 x 15
##   variable    mean     sd  rhat   q2.5     q5    q10    q25    q50    q75    q90
##   <chr>      <dbl>  <dbl> <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1 pop_by_r~ 1.90e5 11848.  1.00 1.68e5 1.71e5 1.76e5 1.82e5 1.90e5 1.98e5 2.05e5
## 2 pop_by_r~ 3.48e6 46048.  1.00 3.40e6 3.41e6 3.43e6 3.45e6 3.48e6 3.52e6 3.54e6
## 3 pop_by_r~ 2.15e7 46846.  1.00 2.14e7 2.14e7 2.15e7 2.15e7 2.15e7 2.16e7 2.16e7
## # ... with 4 more variables: q95 <dbl>, q97.5 <dbl>, ess_bulk <dbl>,
## #   ess_tail <dbl>
```

```
# Final counts at high-risk of statelessness by initial risk level
civ_model_results$summary_table %>%
  filter(str_detect(variable, "counts_by_risk_level"))
```

```
## # A tibble: 3 x 15
##   variable    mean     sd  rhat   q2.5     q5    q10    q25    q50    q75    q90
##   <chr>      <dbl>  <dbl> <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1 counts_b~ 1.55e5 2.43e4  1.00 9.88e4 1.10e5 1.22e5 1.41e5 1.58e5 1.73e5 1.84e5
## 2 counts_b~ 1.99e6 1.62e5  1.00 1.67e6 1.72e6 1.78e6 1.87e6 1.99e6 2.10e6 2.20e6
## 3 counts_b~ 0      0      NA    0      0      0      0      0      0      0
## # ... with 4 more variables: q95 <dbl>, q97.5 <dbl>, ess_bulk <dbl>,
## #   ess_tail <dbl>
```

We can also extract more detailed information about the estimates of stateless persons by birth country, region, and risk level. The following chunk of code extract that information.

```
at_risk <- civ_model_results$summary_table %>%
  filter(str_detect(variable, "p_pred")) %>%
  bind_cols(pred_df %>% bind_rows(pred_df) %>% bind_rows(pred_df)) %>%
  mutate(risk_level = ifelse(str_ends(variable, "1]"), "high", ifelse(str_ends(variable, "2]")
  filter(risk_level!="no risk") %>%
  select(birth_country, region, n, risk_level, mean, q2.5, q97.5) %>%
  rename(prop_stateless = mean,
         prop_lower = q2.5,
         prop_upper = q97.5,
         total_pop = n) %>%
  mutate(pop_stateless = total_pop*prop_stateless,
         pop_lower = total_pop*prop_lower,
         pop_upper = total_pop*prop_upper) %>%
  bind_cols(civ_model_results$summary_table %>%
  filter(str_detect(variable, "unweighted_p")) %>%
  select(variable, mean) %>%
  rename(prob_failure = mean) %>%
  mutate(risk_level = ifelse(str_ends(variable, "1]"), "high", ifelse(str_ends(variable, "2]")
  filter(risk_level!="no risk") %>%
  select(-variable) %>%
  bind_cols(civ_model_results$summary_table %>%
  filter(str_detect(variable, "r_props_pred")) %>%
  select(variable, mean) %>%
  rename(prob_risk = mean) %>%
  mutate(risk_level = ifelse(str_ends(variable, "1]"), "high", ifelse(str_ends(variable, "2]")
  filter(risk_level!="no risk") %>%
  select(prob_risk)) %>%
  select(prob_risk, prob_failure)) %>%
  select(birth_country, region, total_pop, risk_level, prob_risk, prob_failure, prop_stateless
```

We can just look at the total population by birth/region combination, and the estimated population stateless in that combination and risk level. For example, the first row in the output below suggests out of a total population of around 7200, 211 are estimated to be stateless.

```
at_risk %>%
  select(birth_country, region, risk_level, total_pop, pop_stateless)
```

```
## # A tibble: 1,518 x 5
##    birth_country                region        risk_level total_pop pop_stateless
##    <chr>                        <fct>         <chr>          <dbl>         <dbl>
##  1 BENIN                        DISTRICT AUT~ high           7202.         211. 
##  2 BURKINA-FASO                 DISTRICT AUT~ high         123642.         936. 
##  3 BURUNDI                      DISTRICT AUT~ high           1200.          23.3
##  4 CAMEROUN                     DISTRICT AUT~ high              0            0  
##  5 CENTRAFRIQUE                 DISTRICT AUT~ high              0            0  
##  6 CONGO BRAZZAVILLE            DISTRICT AUT~ high           7202.         126. 
##  7 CONGO KINSHASA (RDC)         DISTRICT AUT~ high           3601.          67.7
##  8 COTE D'IVOIRE                DISTRICT AUT~ high        5154569.       22576. 
##  9 ETATS UNIS D'AMERIQUES (USA) DISTRICT AUT~ high           1200.          24.9
## 10 GABON                        DISTRICT AUT~ high              0            0  
## # ... with 1,508 more rows
```

The following code would save this `at_risk` dataframe as a csv.

```
write_csv(at_risk, "results_by_bp_region_risk.csv")
```