Data Section

Nick Callow

10/15/2020

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   caseid = col_double(),
##   age = col_double(),
##   age_first_child = col_double(),
##   age_youngest_child_under_6 = col_double(),
##   total_children = col_double(),
##   age_start_relationship = col_double(),
##   age_at_first_marriage = col_double(),
##   age_at_first_birth = col_double(),
##   distance_between_houses = col_double(),
##   age_youngest_child_returned_work = col_double(),
##   feelings_life = col_double(),
##   hh_size = col_double(),
##   number_total_children_intention = col_double(),
##   number_marriages = col_double(),
##   fin_supp_child_supp = col_double(),
##   fin_supp_child_exp = col_double(),
##   fin_supp_lump = col_double(),
##   fin_supp_other = col_double(),
##   is_male = col_double(),
##   main_activity = col_logical()
##   # ... with 1 more columns
## )
```

```
## See spec(...) for full column specifications.
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──
```

```
## ✓ ggplot2 3.3.2     ✓ dplyr   1.0.2
## ✓ tibble  3.0.3     ✓ stringr 1.4.0
## ✓ tidyr   1.1.2     ✓ forcats 0.5.0
## ✓ purrr   0.3.4
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Data

The data set selected for analysis is the 2017 General Social Survey (GSS). Data collection took place between February and November of 2017, and the goal of the survey is to learn more about Canadian family structures (Statistics Canada, 2017). In this section, we provide an analysis of the survey, including its key features, methdology, and merits.

Target Population and Sampling Frame

To learn more about Canadian family structures, the GSS focuses on a specific section of Canadians. This group, called the *target population* is “all non-institutionalized persons 15 years of age or older, living in the ten provinces of Canada” (Statistics Canada, 2017, Data Sources and Methdology Section). Of note is that residents of the territories are excluded from the GSS. This decision by Statistics Canada is explored further in the *Weaknesses* section below.

To survey individuals in the target population, we need to know who they are and how to contact them. This list of people is called the *sampling frame*, and Statistics Canada recruits individuals from this list (according to the sampling approach detailed below). The sampling frame of the GSS is built in the following way. First, a list of landline and cellular phone numbers are obtained from the Census (Statistics Canada, 2017). These phone numbers are then grouped into households if they share the same address (Statistics Canada, 2017). Ideally, everyone in the target population is included in the sampling frame (Wu & Thompson, 2020). However, when this is the not the case, it introduces a bias in the survey known as *coverage error*. The sampling frame of the GSS does not capture everyone in the target population. Specifically, individuals without telephone service, or a known phone number, will be excluded (Statistics Canada, 2017). This coverage error implies that the GSS may not be entirely representative of the Canadian population, and therefore, our estimates should be approached cautiously.

Sampling Approach

After creating the sampling frame, Statistics Canada needs to decide how to pick individuals from this list, since it would be too costly to survey them all. The method of selecting Canadians from the frame is called the *sampling approach*. The GSS employs a *stratified random sampling approach*. First, the target population is divided into smaller groups, called strata, where individuals share a common characteristics or property (Caetano, 2020b). In this survey, Canadians are grouped by geographic location for a total of twenty seven strata (Statistics Canada, 2017). Second, a percentage of households in each stratum are randomly selected (Statistics Canada, 2017). Third, one eligible member of each chosen household is randomly selected for participation (Statistics Canada, 2017). These last two steps employ *simple random sampling*, meaning survey participants are picked randomly, and each member of the sampling frame has an equal chance of selection (Caetano, 2020b).

There are several benefits to this method. First, the GSS uses a probability sampling approach (stratified random sampling). The odds of any one individual being included in the survey are known, and can be weighted to reflect the characteristics of the populations (Caetano, 2020a). Therefore, we know that our data reflects well Canadian family structures. Second, this approach can be replicated, allowing for verification and comparison (Caetano, 2020a). However, there are some limitations. First, probability sampling, including stratified random sampling, is very costly. The decline in GSS sample sizes since 1999 may reflect this fact (Statistics Canada, 2019). Second, random selection implies that participants may not want to participate in the survey (Caetano, 2020a). Therefore, the non-response rate may be high, and this may introduce bias into the survey (see Sampled Population and Missing data below).

Sampled Population and Missing Data

Not everyone selected to participate in the survey will do so. The set of individuals who take part are called the *sampled population* (Wu & Thompson, 2020). In the GSS, a total of $n = 20,602$ individuals completed some or all of the survey questions. This yieled a response rate of 52.6%.

Missing data is a concern for any survey. Individuals who agree to participate may not complete all the questions. When this happens, holes appear in the data, and this makes analysis more difficult. As seen in Figure 1, all of our predictor variables have limited missing data. In contrast, our response variable, *age_at_first_birth*, has a non-response rate of 38.18%. While this may seem large and concerning, Figure 2 clarifies that this rate is inflated. Individuals without any children (that is, $\text{total\_children} = 0$), make up the largest proportion of the non-responses. Therefore, we consider the variable still worthy and capable of analysis.

Figure 1: Missing Data in Variables of Interest

References

Caetano, S. (2020a, September 21). Probability versus Non-Probability Sampling [PowerPoint Slides]. Quercus. https://q.utoronto.ca/courses/184060/files/8975319?module_item_id=1816946 (https://q.utoronto.ca/courses/184060/files/8975319?module_item_id=1816946)

Caetano, S. (2020b, September 28). Sampling Techniques [PowerPoint Slides]. Quercus. https://q.utoronto.ca/courses/184060/files/9058287?module_item_id=1828122 (https://q.utoronto.ca/courses/184060/files/9058287?module_item_id=1828122)

Statistics Canada (2017, December 19). *General Social Survey - Family (GSS)*. Government of Canada. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=335816 (https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=335816)

Wu, C., & Thompson, M. E. (2020). *Sampling Theory and Practice.* Cham: Springer International Publishing.