R Users Breakout Session

Matt Gunther - IPUMS PMA Senior Data Analyst

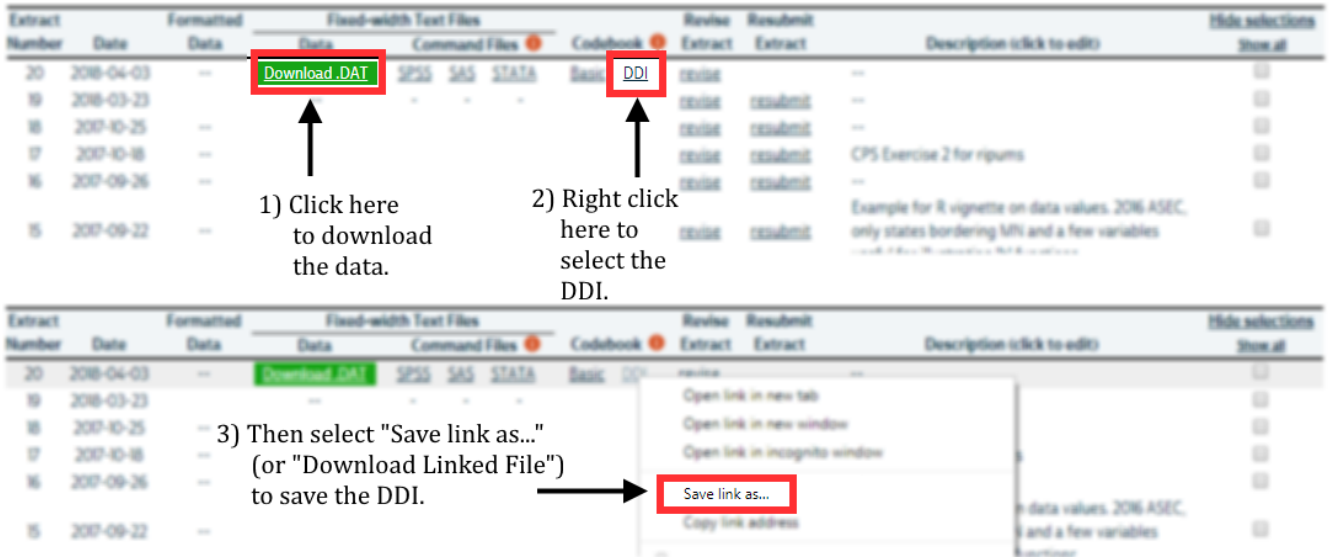
2022-04-06

Setup

R users: remember to select a **.dat (fixed-width text)** data format

You'll receive a compressed **dat.gz** file - no need to decompress!

Save both of those files in the "data" folder of your working directory.



You'll need the ipumsr package to load them. If not installed, you can download from CRAN.

```
install.packages("ipumsr")
```

Each session, load the `ipumsr` library before you import data.

```
library(ipumsr)

# Load data into R with `ipumsr`
dat <- read_ipums_micro(
  ddi = "data/pma_00093.xml",
  data = "data/pma_00093.dat.gz"
)
```

Other useful packages for IPUMS data:

```
# General toolkit
library(tidyverse)

# For label manipulation:
library(labelled)

# For survey analysis:
library(survey)
library(srvyr)
```

1 - Analytic Sample

PMA uses an **open panel design** - women may enter the panel after Phase 1, and they may be lost to follow-up after any phase.

See RESULTFQ

Women who enter the panel at Phase 2 are `NA` for all variables at Phase 1.

```
dat %>% count(RESULTFQ_1)
```

```
## # A tibble: 3 × 2
##               RESULTFQ_1     n
##                <int+lbl> <int>
## 1  1 [Completed]        16314
## 2  5 [Partly completed]    34
## 3 NA                     4514
```

Women whose households were not found again after Phase 1 are `NA` for all variables at Phase 2.

```
dat %>% count(RESULTFQ_2)
```

```
## # A tibble: 11 × 2
##                                       RESULTFQ_2     n
##                                        <int+lbl> <int>
##  1  1 [Completed]                                17015
##  2  2 [Not at home]                                106
##  3  3 [Postponed]                                   24
##  4  4 [Refused]                                     87
##  5  5 [Partly completed]                            22
##  6  7 [Respondent moved]                            18
##  7 10 [Incapacitated]                               24
##  8 95 [Not interviewed (female questionnaire)]       4
##  9 96 [Not interviewed (household questionnaire)]  197
## 10 99 [NIU (not in universe)]                     1353
## 11 NA                                             2012
```

We will only include women who were available and completed the Female Questionnaire for *both* Phase 1 and Phase 2.

```
dat <- dat %>% filter(RESULTFQ_1 == 1 & RESULTFQ_2 == 1)

dat %>% count(RESULTFQ_1, RESULTFQ_2)
```

```
## # A tibble: 1 × 3
##      RESULTFQ_1    RESULTFQ_2     n
##       <int+lbl>     <int+lbl> <int>
## 1 1 [Completed] 1 [Completed] 12501
```

Additionally, PMA samples are only valid for the *de facto* population: women who slept in the household the night before the Household interview.

See RESIDENT

```
dat %>% count(RESIDENT_1)
```

```
## # A tibble: 3 × 2
##                                          RESIDENT_1     n
##                                           <int+lbl> <int>
## 1 11 [Visitor, slept in hh last night]                140
## 2 21 [Usual member, did not sleep in hh last night]   194
## 3 22 [Usual member, slept in hh last night]         12167
```

We'll also drop cases where the woman was not part of the *de facto* population in either Phase 1 or Phase 2.

```
dat <- dat %>% filter(RESIDENT_1 %in% c(11, 22) & RESIDENT_2 %in% c(11, 22))
```

How many cases remain?

```
dat %>% count(COUNTRY)
```

```
## # A tibble: 2 × 2
##            COUNTRY     n
##          <int+lbl> <int>
## 1 1 [Burkina Faso]  5208
## 2 7 [Kenya]         6935
```

2 - Recoding Independent variables

PMA surveys contain many **categorical** variables. These are usually represented as **factors** in R.

In an IPUMS data extract, you won't see factors!

Instead, we generate **labelled** numeric variables (note the label in brackets).

```
dat %>% ipums_var_label(CVINCOMELOSS_2)
```

```
## [1] "Income loss resulted from COVID-19 restrictions"
```

```
dat %>% count(CVINCOMELOSS_2)
```

```
## # A tibble: 4 × 2
##               CVINCOMELOSS_2     n
##                    <int+lbl> <int>
## 1  0 [No]                        658
## 2  1 [Yes]                      7566
## 3 97 [Don't know]                  2
## 4 99 [NIU (not in universe)]    3917
```

The ipumsr package contains tools for working with labelled IPUMS data.

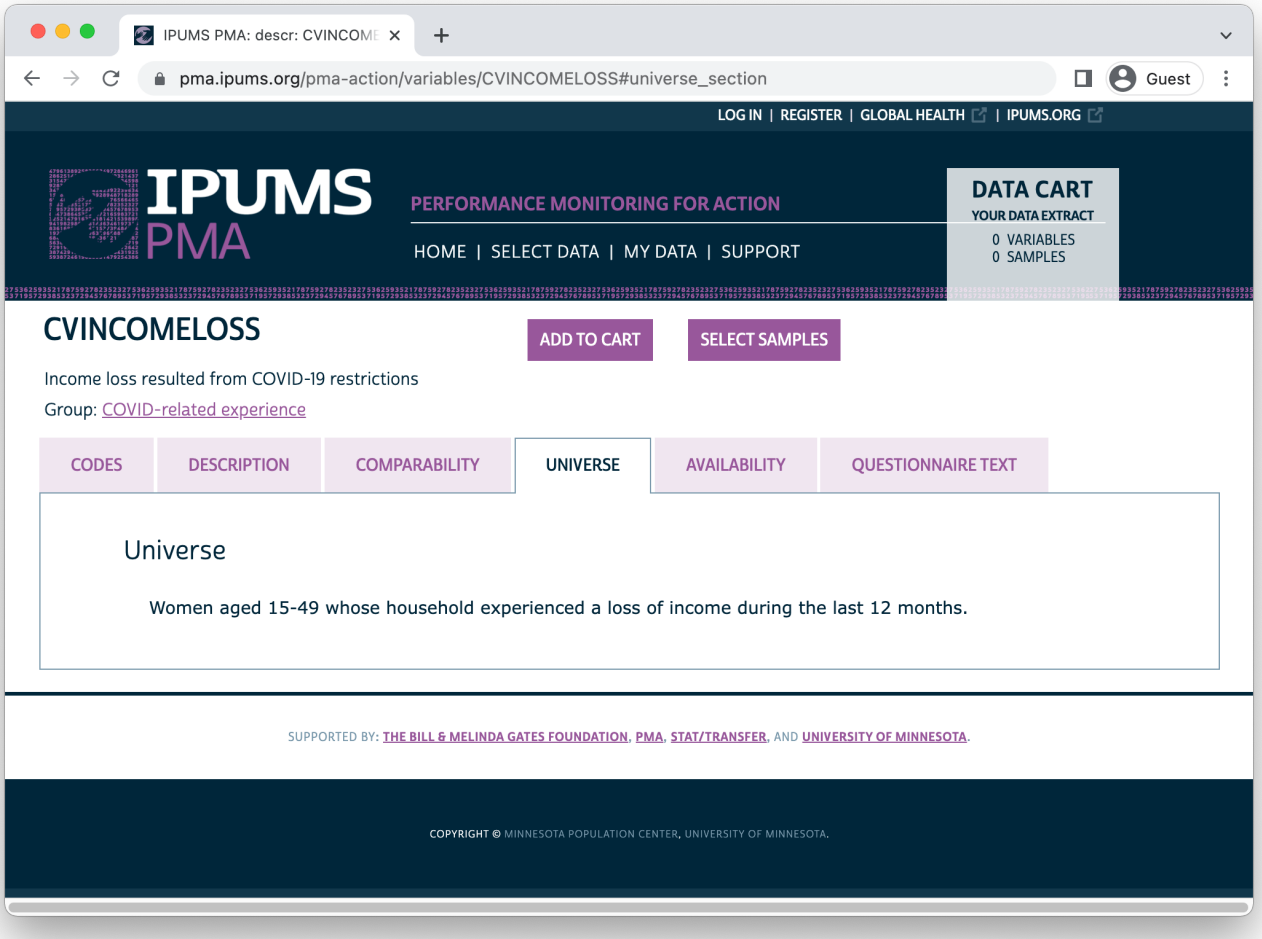
Usually, we handle codes like `99 [NIU (not in universe)]` before transforming other missing data to `NA`.

```
dat %>% count(CVINCOMELOSS_2, HHINCOMELOSSAMT_2)
```

```
## # A tibble: 7 × 3
##               CVINCOMELOSS_2           HHINCOMELOSSAMT_2     n
##                    <int+lbl>                    <int+lbl> <int>
## 1  0 [No]                     2 [Partial]                  547
## 2  0 [No]                     3 [Complete]                 111
## 3  1 [Yes]                    2 [Partial]                 5449
## 4  1 [Yes]                    3 [Complete]                2117
## 5 97 [Don't know]             2 [Partial]                    2
## 6 99 [NIU (not in universe)]  1 [None]                    3904
## 7 99 [NIU (not in universe)] 98 [No response or missing]    13
```

Tip:

Information the code `NIU (not in universe)` can always be found on a variable's universe tab.



For CVINCOMELOSS_2, `99 [NIU (not in universe)]` may indicate that the household experienced *no income loss in the last year*, or it may indicate that HHINCOMELOSSAMT_2 is `98 [No response or missing]`.

We should treat the `NIU` women from households without *any* income loss as "No" in CVINCOMELOSS_2.

```
dat <- dat %>%
  mutate(
    CVINCOMELOSS_2 = CVINCOMELOSS_2 %>%
      labelled::recode_if(HHINCOMELOSSAMT_2 == 1, 0)
  )

dat %>% count(CVINCOMELOSS_2, HHINCOMELOSSAMT_2)
```

```
## # A tibble: 7 × 3
##               CVINCOMELOSS_2           HHINCOMELOSSAMT_2     n
##                    <int+lbl>                   <int+lbl> <int>
## 1  0 [No]                     1 [None]                     3904
## 2  0 [No]                     2 [Partial]                   547
## 3  0 [No]                     3 [Complete]                  111
## 4  1 [Yes]                    2 [Partial]                  5449
## 5  1 [Yes]                    3 [Complete]                 2117
## 6 97 [Don't know]             2 [Partial]                     2
## 7 99 [NIU (not in universe)] 98 [No response or missing]    13
```

Next, we'll use `NA` to represent the remaining values above `90`:

- `97 [Don't know]` and
- remaining cases marked `99 [NIU (not in universe)]`

```
dat <- dat %>%
  mutate(
    CVINCOMELOSS_2 = CVINCOMELOSS_2 %>%
      lbl_na_if(~.val > 90)
  )

dat %>% count(CVINCOMELOSS_2, HHINCOMELOSSAMT_2)
```

```
## # A tibble: 7 × 3
##   CVINCOMELOSS_2           HHINCOMELOSSAMT_2     n
##        <int+lbl>                   <int+lbl> <int>
## 1        0 [No]   1 [None]                    3904
## 2        0 [No]   2 [Partial]                  547
## 3        0 [No]   3 [Complete]                 111
## 4        1 [Yes]  2 [Partial]                5449
## 5        1 [Yes]  3 [Complete]               2117
## 6       NA        2 [Partial]                   2
## 7       NA       98 [No response or missing]    13
```

Once you're done with labels, we recommend transforming key variables into **factors** with forcats::as_factor.

The forcats package is included when you load `library(tidyverse)`.

```
dat <- dat %>% mutate(CVINCOMELOSS_2 = as_factor(CVINCOMELOSS_2))

dat %>% count(CVINCOMELOSS_2)
```

```
## # A tibble: 3 × 2
##   CVINCOMELOSS_2     n
##   <fct>          <int>
## 1 No              4562
## 2 Yes             7566
## 3 <NA>              15
```

This will make categorical variables easier to use in data visualization and as "dummy" variables in regression analysis.

Likert-style questions can be treated as factors, too.

```
dat %>% ipums_var_label(COVIDCONCERN_2)
```

```
## [1] "Concerned about getting infected"
```

```
dat %>% count(COVIDCONCERN_2)
```

```
## # A tibble: 6 × 2
##                                       COVIDCONCERN_2     n
##                                            <int+lbl> <int>
## 1  1 [Not concerned]                                    374
## 2  2 [A little concerned]                               677
## 3  3 [Concerned]                                       2470
## 4  4 [Very concerned]                                  8610
## 5  5 [Currently / previously infected with COVID-19]      9
## 6 98 [No response or missing]                            3
```

This time we'll treat codes `5` and above as `NA`.

```
dat <- dat %>%
  mutate(
    COVIDCONCERN_2 = COVIDCONCERN_2 %>%
      lbl_na_if(~.val >= 5) %>%
      as_factor()
  )

dat %>% count(COVIDCONCERN_2)
```

```
## # A tibble: 5 × 2
##   COVIDCONCERN_2          n
##   <fct>               <int>
## 1 Not concerned         374
## 2 A little concerned    677
## 3 Concerned            2470
## 4 Very concerned       8610
## 5 <NA>                   12
```

You can apply the same transformation to several variables with help from dplyr::across.

dplyr is another package included when you load `library(tidyverse)`.

```
dat <- dat %>%
  mutate(
    across(
      c(COUNTRY, URBAN, WEALTHT_2, EDUCATTGEN_2),
      ~.x %>% lbl_na_if(~.val >= 90) %>% as_factor()
    )
  )
```

Often, it's important to set a **reference group** against which all dummy variables will be compared.

You can manually specify a **refernece group** when you set factor "levels" with a function like forcats::fct_relevel.

```
dat <- dat %>%
  mutate(
    AGE_2 = case_when(
      AGE_2 < 25 ~ "15-24",
      AGE_2 < 35 ~ "25-34",
      AGE_2 < 50 ~ "35-49"
    ),
    AGE_2 = AGE_2 %>% fct_relevel("15-24", "25-34", "35-49")
  )
```

3 - Dependent variables

We'll use our recoded variables to model the likelihood of contraceptive method **adoption** and **discontinuation** between phases.

See CP

```
dat <- dat %>% filter(CP_1 < 90 & CP_2 < 90)

dat %>% count(CP_1, CP_2)
```

```
## # A tibble: 4 × 3
##        CP_1      CP_2     n
##   <int+lbl> <int+lbl> <int>
## 1   0 [No]    0 [No]   5107
## 2   0 [No]    1 [Yes]  1939
## 3   1 [Yes]   0 [No]   1178
## 4   1 [Yes]   1 [Yes]  3917
```

A woman has **adopted** a method if she was *not* using one at Phase 1, but then reported using one at Phase 2.

She has **discontinued** a method if she *did* use one at Phase 1, but no longer uses one at Phase 2.

```
dat <- dat %>%
  mutate(
    FPSTATUS = case_when(
      CP_1 == 1 & CP_2 == 1 ~ "User",
      CP_1 == 0 & CP_2 == 0 ~ "Non-user",
      CP_1 == 1 & CP_2 == 0 ~ "Discontinued",
      CP_1 == 0 & CP_2 == 1 ~ "Adopted"
    ),
    FPSTATUS = fct_infreq(FPSTATUS)
  )
```

Un-weighted sample proportions for `FPSTATUS` can be found with count and prop.table:

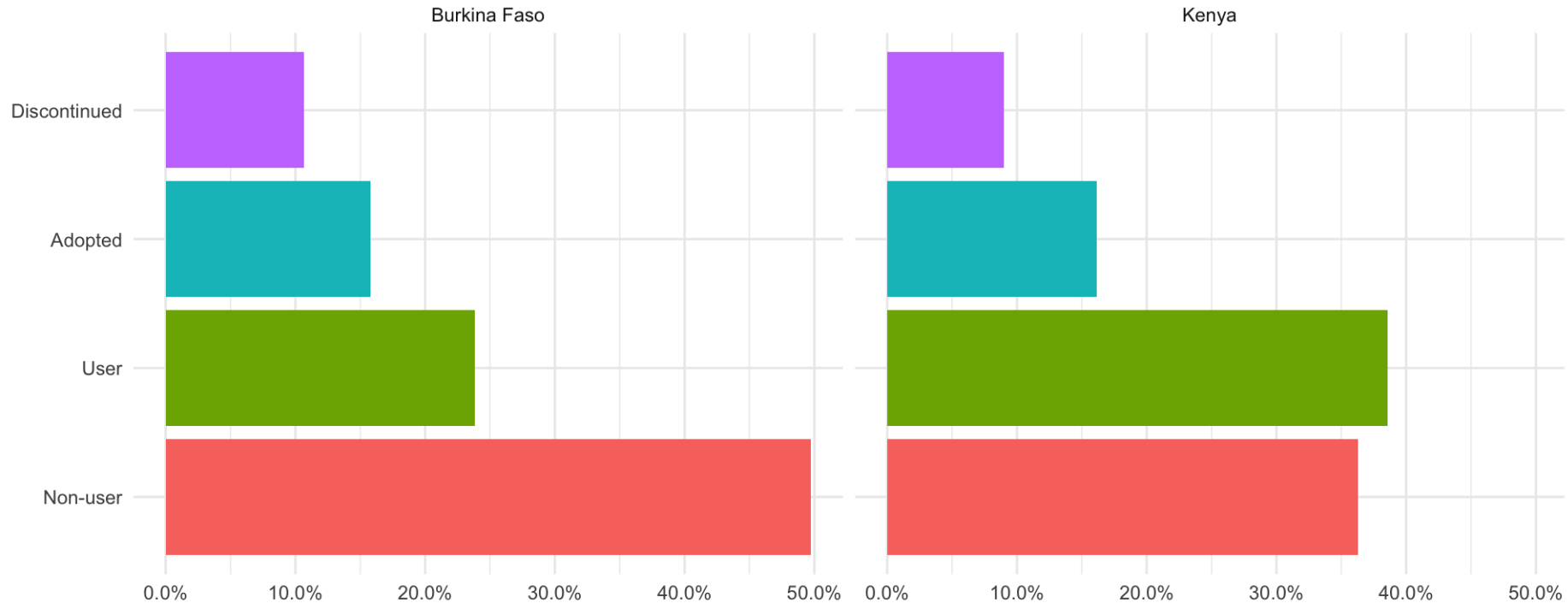
```
dat_nowt <- dat %>%
  group_by(COUNTRY) %>%
  count(FPSTATUS) %>%
  mutate(prop = prop.table(n))

dat_nowt
```

```
## # A tibble: 8 × 4
## # Groups:   COUNTRY [2]
##   COUNTRY      FPSTATUS         n   prop
##   <fct>        <fct>        <int>  <dbl>
## 1 Burkina Faso Non-user      2589 0.497
## 2 Burkina Faso User          1241 0.238
## 3 Burkina Faso Adopted        821 0.158
## 4 Burkina Faso Discontinued   556 0.107
## 5 Kenya        Non-user      2518 0.363
## 6 Kenya        User          2676 0.386
## 7 Kenya        Adopted       1118 0.161
## 8 Kenya        Discontinued   622 0.0897
```

We'll plot this table with ggplot2 (also included with the tidyverse).

```
dat_nowt %>%
  ggplot(aes(x = prop, y = FPSTATUS, fill = FPSTATUS)) +
  geom_bar(stat = "identity") +
  facet_wrap(~COUNTRY) + theme_minimal() +
  theme(axis.title = element_blank(), legend.position = "none") +
  scale_x_continuous(labels = scales::label_percent())
```



For *weighted* population estimates, use as_survey_design and survey_mean from the srvyr package.

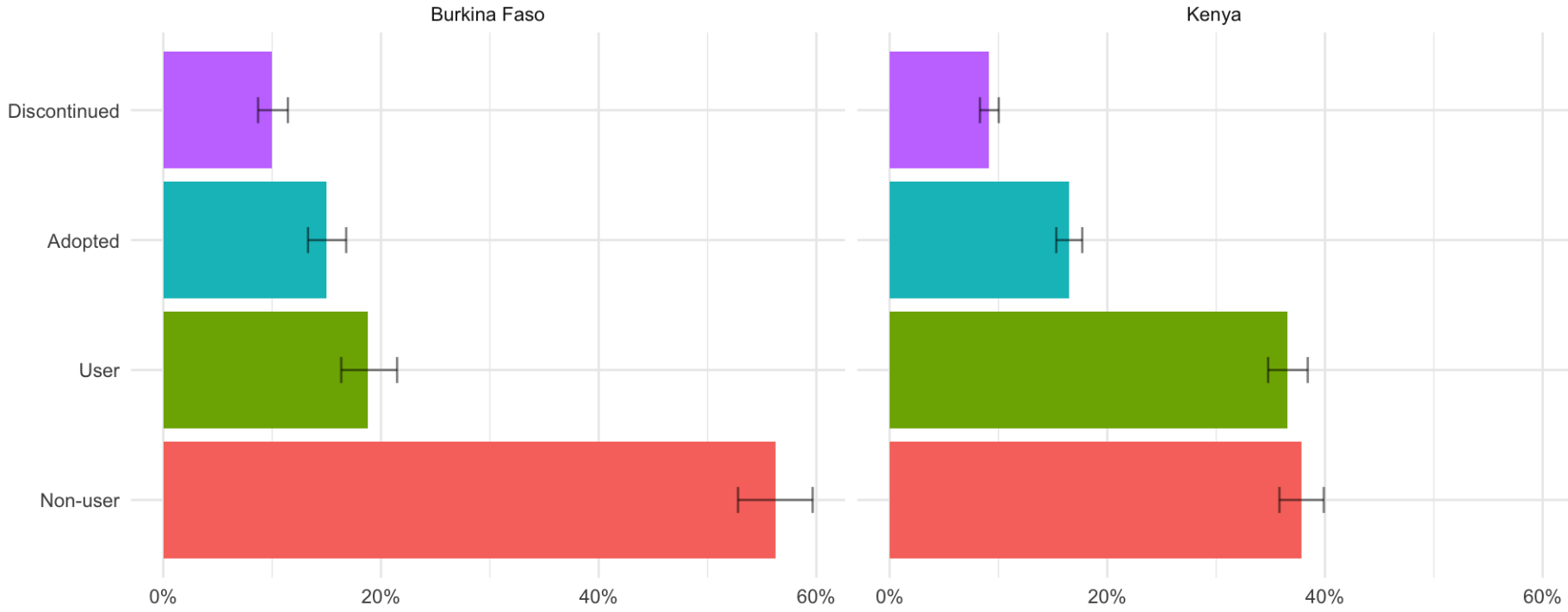
Use `prop = TRUE` to adjust standard errors near 0% or 100% for proportions.

```
dat_wtd <- dat %>%
  as_survey_design(weight = PANELWEIGHT, id = EAID_1, strata = COUNTRY) %>%
  group_by(COUNTRY, FPSTATUS) %>%
  summarise(survey_mean(prop = TRUE, prop_method = "logit", vartype = "ci"))

dat_wtd
```

```
## # A tibble: 8 × 5
## # Groups:   COUNTRY [2]
##   COUNTRY      FPSTATUS       coef `_low` `_upp`
##   <fct>        <fct>         <dbl>  <dbl>  <dbl>
## 1 Burkina Faso Non-user     0.563  0.528   0.597
## 2 Burkina Faso User         0.188  0.163   0.215
## 3 Burkina Faso Adopted      0.150  0.133   0.168
## 4 Burkina Faso Discontinued 0.0999 0.0871  0.114
## 5 Kenya        Non-user     0.378  0.358   0.399
## 6 Kenya        User         0.366  0.348   0.384
## 7 Kenya        Adopted      0.165  0.153   0.177
## 8 Kenya        Discontinued 0.0912 0.0830  0.100
```

```
dat_wtd %>%
  ggplot(aes(x = coef, y = FPSTATUS, fill = FPSTATUS)) +
  geom_bar(stat = "identity") +
  geom_errorbar(aes(xmin = `_low`, xmax = `_upp`), width = 0.2, alpha = 0.5) +
  facet_wrap(~COUNTRY) + theme_minimal() +
  theme(axis.title = element_blank(), legend.position = "none") +
  scale_x_continuous(labels = scales::label_percent())
```



4 - Analysis

The same srvyr toolkit can be used to model our dependent variables with survey::svyglm.

Consider women who were *not* using a method at Phase 1:

```
adopt_glm <- dat %>%
  filter(CP_1 == 0) %>%
  mutate(adopt = FPSTATUS == "Adopted") %>%
  group_by(COUNTRY) %>%
  summarise(
    adopt = cur_data() %>%
      as_survey_design(weight = PANELWEIGHT, id = EAID_1, strata = STRATA_1) %>%
      svyglm(
        adopt ~ CVINCOMELOSS_2 + COVIDCONCERN_2 + URBAN + WEALTHT_2 + EDUCATTGEN_2 + AGE_2,
        family = "quasibinomial", design = .
      ) %>%
      broom::tidy(exp = TRUE) %>%
      mutate(sig = gtools::stars.pval(p.value)) %>%
      list()
  )

adopt_glm
```

```
## # A tibble: 2 × 2
##   COUNTRY      adopt
##   <fct>        <list>
## 1 Burkina Faso <tibble [13 × 6]>
## 2 Kenya        <tibble [13 × 6]>
```

For Phase 1 non-users in Burkina Faso, **very high** levels of concern about becoming infected with COVID-19 are significantly associated with higher chances of adopting a contraceptive method (relative to women who had no such concern).

Lesser levels of concern are not statistically significant, nor is household income loss from COVID-19.

```
adopt_glm %>%
  filter(COUNTRY == "Burkina Faso") %>%
  unnest(adopt)
```

```
## # A tibble: 13 × 7
##    COUNTRY      term                                 estimate std.error statistic       p.value sig
##    <fct>        <chr>                                   <dbl>     <dbl>     <dbl>         <dbl> <chr>
##  1 Burkina Faso (Intercept)                           0.0985     0.366    -6.33  0.00000000262 "***"
##  2 Burkina Faso CVINCOMELOSS_2Yes                     1.28       0.155     1.61  0.109         " "
##  3 Burkina Faso COVIDCONCERN_2A little concerned      1.80       0.373     1.58  0.117         " "
##  4 Burkina Faso COVIDCONCERN_2Concerned               1.37       0.351     0.891 0.375         " "
##  5 Burkina Faso COVIDCONCERN_2Very concerned          1.91       0.318     2.02  0.0446        "*"
##  6 Burkina Faso URBANUrban                            1.36       0.186     1.65  0.101         " "
##  7 Burkina Faso WEALTHT_2Middle tertile               0.962      0.170    -0.230 0.818         " "
##  8 Burkina Faso WEALTHT_2Highest tertile              0.735      0.220    -1.40  0.164         " "
##  9 Burkina Faso EDUCATTGEN_2Primary/Middle school     1.44       0.161     2.24  0.0265        "*"
## 10 Burkina Faso EDUCATTGEN_2Secondary/post-primary    1.51       0.181     2.27  0.0246        "*"
## 11 Burkina Faso EDUCATTGEN_2Tertiary/post-secondary   2.30       0.352     2.37  0.0192        "*"
## 12 Burkina Faso AGE_225-34                            1.72       0.180     3.02  0.00298       "**"
## 13 Burkina Faso AGE_235-49                            1.08       0.195     0.385 0.701         " "
```

In Kenya, neither of these measures are significantly predictive of adoption among non-users.

```
adopt_glm %>%
  filter(COUNTRY == "Kenya") %>%
  unnest(adopt)
```

```
## # A tibble: 13 × 7
##    COUNTRY term                                estimate std.error statistic  p.value sig
##    <fct>   <chr>                                  <dbl>     <dbl>     <dbl>    <dbl> <chr>
##  1 Kenya   (Intercept)                            0.104     0.371    -6.09  3.76e- 9 "***"
##  2 Kenya   CVINCOMELOSS_2Yes                      1.20      0.111     1.61  1.08e- 1 " "
##  3 Kenya   COVIDCONCERN_2A little concerned       0.645     0.351    -1.25  2.13e- 1 " "
##  4 Kenya   COVIDCONCERN_2Concerned                0.794     0.256    -0.900 3.69e- 1 " "
##  5 Kenya   COVIDCONCERN_2Very concerned           0.907     0.254    -0.385 7.00e- 1 " "
##  6 Kenya   URBANUrban                             1.17      0.147     1.06  2.92e- 1 " "
##  7 Kenya   WEALTHT_2Middle tertile                1.12      0.112     1.01  3.15e- 1 " "
##  8 Kenya   WEALTHT_2Highest tertile               0.817     0.151    -1.34  1.80e- 1 " "
##  9 Kenya   EDUCATTGEN_2Primary/Middle school      2.30      0.273     3.05  2.53e- 3 "**"
## 10 Kenya   EDUCATTGEN_2Secondary/post-primary     2.87      0.302     3.49  5.54e- 4 "***"
## 11 Kenya   EDUCATTGEN_2Tertiary/post-secondary    3.63      0.306     4.21  3.51e- 5 "***"
## 12 Kenya   AGE_225-34                             3.06      0.128     8.71  2.98e-16 "***"
## 13 Kenya   AGE_235-49                             1.61      0.131     3.62  3.53e- 4 "***"
```

What about method **dicontinuation** for women who *were* using a method at Phase 1?

```
stop_glm <- dat %>%
  filter(CP_1 == 1) %>%
  mutate(stop = FPSTATUS == "Discontinued") %>%
  group_by(COUNTRY) %>%
  summarise(
    stop = cur_data() %>%
      as_survey_design(weight = PANELWEIGHT, id = EAID_1, strata = STRATA_1) %>%
      svyglm(
        stop ~ CVINCOMELOSS_2 + COVIDCONCERN_2 + URBAN + WEALTHT_2 + EDUCATTGEN_2 + AGE_2,
        family = "quasibinomial", design = .
      ) %>%
      broom::tidy(exp = TRUE) %>%
      mutate(sig = gtools::stars.pval(p.value)) %>%
      list()
  )

stop_glm
```

```
## # A tibble: 2 × 2
##   COUNTRY      stop
##   <fct>        <list>
## 1 Burkina Faso <tibble [13 × 6]>
## 2 Kenya        <tibble [13 × 6]>
```

This time, neither of the COVID-19 measures are significantly associated with **discontinuation** for Phase 1 contraceptive users in Burkina Faso.

```
stop_glm %>%
  filter(COUNTRY == "Burkina Faso") %>%
  unnest(stop)
```

```
## # A tibble: 13 × 7
##    COUNTRY      term                                 estimate std.error statistic p.value sig
##    <fct>        <chr>                                   <dbl>     <dbl>     <dbl>   <dbl> <chr>
##  1 Burkina Faso (Intercept)                             0.536     0.407    -1.53   0.127  " "
##  2 Burkina Faso CVINCOMELOSS_2Yes                       0.857     0.185    -0.835  0.405  " "
##  3 Burkina Faso COVIDCONCERN_2A little concerned        1.18      0.442     0.379  0.705  " "
##  4 Burkina Faso COVIDCONCERN_2Concerned                 0.922     0.425    -0.192  0.848  " "
##  5 Burkina Faso COVIDCONCERN_2Very concerned            0.935     0.335    -0.200  0.842  " "
##  6 Burkina Faso URBANUrban                              0.951     0.231    -0.215  0.830  " "
##  7 Burkina Faso WEALTHT_2Middle tertile                 1.47      0.211     1.82   0.0702 "."
##  8 Burkina Faso WEALTHT_2Highest tertile                0.797     0.238    -0.952  0.343  " "
##  9 Burkina Faso EDUCATTGEN_2Primary/Middle school       1.29      0.212     1.21   0.226  " "
## 10 Burkina Faso EDUCATTGEN_2Secondary/post-primary      1.16      0.250     0.596  0.552  " "
## 11 Burkina Faso EDUCATTGEN_2Tertiary/post-secondary     0.787     0.289    -0.828  0.409  " "
## 12 Burkina Faso AGE_225-34                              1.11      0.215     0.482  0.630  " "
## 13 Burkina Faso AGE_235-49                              0.784     0.244    -0.997  0.320  " "
```

However, higher levels concern with becoming infected with COVID-19 *are* significantly associated with higher odds of discontinuation for Phase 1 contraceptive users in Kenya.

```
stop_glm %>%
  filter(COUNTRY == "Kenya") %>%
  unnest(stop)
```

```
## # A tibble: 13 × 7
##    COUNTRY term                                estimate std.error statistic  p.value sig
##    <fct>   <chr>                                  <dbl>     <dbl>     <dbl>    <dbl> <chr>
##  1 Kenya   (Intercept)                           0.0978     0.877   -2.65   0.00853  "**"
##  2 Kenya   CVINCOMELOSS_2Yes                     1.01       0.158    0.0433 0.965    " "
##  3 Kenya   COVIDCONCERN_2A little concerned      7.68       0.694    2.94   0.00360  "**"
##  4 Kenya   COVIDCONCERN_2Concerned               4.24       0.723    2.00   0.0467   "*"
##  5 Kenya   COVIDCONCERN_2Very concerned          3.77       0.719    1.85   0.0661   "."
##  6 Kenya   URBANUrban                            1.12       0.135    0.836  0.404    " "
##  7 Kenya   WEALTHT_2Middle tertile               0.843      0.153   -1.11   0.266    " "
##  8 Kenya   WEALTHT_2Highest tertile              0.888      0.180   -0.659  0.511    " "
##  9 Kenya   EDUCATTGEN_2Primary/Middle school     0.787      0.349   -0.687  0.493    " "
## 10 Kenya   EDUCATTGEN_2Secondary/post-primary     0.958      0.367   -0.118  0.907    " "
## 11 Kenya   EDUCATTGEN_2Tertiary/post-secondary    1.10       0.397    0.238  0.812    " "
## 12 Kenya   AGE_225-34                            0.783      0.153   -1.60   0.110    " "
## 13 Kenya   AGE_235-49                            0.589      0.153   -3.45   0.000651 "***"
```

For more R tips for IPUMS data, check out:

- The IPUMS PMA blog
- The ipumsr documentation website
- The ipums tutorials page

Thank you!