Metodología para Encuestas

18/05/2015

Índice

1. Introducción

Este documento detalla el método de regresión multinivel y postestratificación (MRP por sus siglas en inglés) para poder corregir sesgo de selección en los datos recolectados a través de una encuesta con diseño muestral no probabilístico por cuotas, como las realizadas en los diagnósticos participativos de Morelos de 2013 y 2014. El objetivo es construir estimadores precisos sobre la población objetivo.

Existen varios factores que pueden afectar o no la calidad de los datos recolectados en una encuesta. La recolección de encuestas realizadas en Morelos en 2013 y 2014 se realizó con las siguientes características: a) muestreo no probabilístico por cuotas con inferencia basada en modelos y no en diseño, b) administración remota del proceso de la encuesta o trabajo de campo, c) recolección de datos asistida por tabletas móviles[1].

Fricker (2008, p. 198) desglosa cuatro posibles fuentes de error en encuestas:

1. **Error de cobertura**: parte de la población no puede ser incluida en la muestra. Para reducir el error de cobertura, las técnicas usadas comúnmente son tres: a) especificar un marco de muestreo tan completo como sea posible, b) utilizar una estrategia de muestreo sin marco en donde casi toda la población tenga una probabilidad de ser elemento de la muestra, c) post-estratificación: ponderar la muestra para que sea equivalente a la población de inferencia en ciertas características.

2. **Error de muestreo**: los resultados de diferentes muestras serán distintos. En una encuesta con muestreo aleatorio, el error de muestreo decrece cuando se incrementa el tamaño de muestra[2].

3. **Errores por no respuesta**: ocurren cuando no se recolectan los datos de una unidad muestral completa o de respuestas parciales. La tasa de no respuesta, la razón entre el número de encuestados y el tamaño de muestra, es una medida de cuán posible es generalizar los resultados de una encuesta pues hay menor sesgo de no respuesta.

4. **Errores de medición**: la respuesta de la encuesta difiere de la "respuesta verdadera". Por ejemplo, cuando el encuestado no responde honestamente, cuando se interpreta mal la pregunta o se cometen errores al contestar, al hacer la pregunta o al transcribirla.

La estructura de este documento es la siguiente: en la sección dos se especifican los tipos de muestreo y se profundiza sobre las ventajas y desventajas que ofrecen las encuestas de diseño no probabilístico. Con este tipo de datos, el problema más importante es el sesgo de selección que puede tener una muestra. En la sección tres se enlistan varios métodos que ayudan a corregir este error y en la sección cuatro se profundiza en particular sobre la regresión multinivel y postestratificación para este caso en particular. En la sección cinco se muestra un ejemplo práctico del uso de este método para las encuestas de Morelos de 2013 y 2014.

2. Muestreo

Una buena muestra es intuitivamente, aquella que es representativa de la población de la que fue extraida, es decir, que los resultados de los datos recolectados de ésta son consistentes con los resultados que hubiésemos obtenido si se hubiera censado a la población (Fricker, 2008, p. 197).

El muestreo de encuestas puede ser agrupado en dos categorías: muestreo *probabilístico* y *no probabilístico*. De la definición de muestreo probabilístico sabemos que cada individuo en la población tiene una probabilidad

[1] En particular, la recolección de información de encuestas asistida por computadoras tiene la ventaja de que las respuestas son capturadas inmediatamente de las personas encuestadas y guardadas en una base de datos para su procesamiento futuro. Esto reduce tiempos, costos y errores de transcripción. Facilita, además, flujos complicados en las respuestas del cuestionario, filtros, control de la validez de las respuestas, inclusión de elementos multimedia, entre otros (Vehovar and Manfreda, 2008, p. 179). Así mismo, permite tener un mejor manejo del trabajo de campo pues se pueden ir examinando tanto el avance como los resultados de manera paralela al levantamiento.

[2] Nótese que esto es cierto debido a que, si la muestra es aleatoria, entonces el estimador tiende asintóticamente al verdadero valor del parámetro. Sin embargo, si no es aleatoria, no hay control en los sesgos y esta propiedad no está garantizada.

positiva y conocida de ser seleccionado[3]. En el segundo grupo se incluye, por ejemplo, el muestreo propositivo (Little, 2014, p. 415). En este tipo de muestreos la probabilidad de inclusión de cada unidad no puede determinarse o es decisión de los individuos el participar o no en la encuesta (*opt-in*).

Una de las formas más comunes de muestreo no probabilístico es el muestreo por cuotas. Las unidades son elegidas tal que la muestra posea la misma distribución que una característica conocida de la población. Por ejemplo, si se conoce la distribución de edad y género en la población, se elige una muestra equivalente a esta distribución. En este ejemplo, los entrevistadores reciben una cuota por cada grupo de edad y género y entrevistan a individuos hasta que la cuota se cumple.

El muestreo por cuotas es similar al muestreo estratificado pues agrupa unidades similares. Sin embargo, difieren en la manera en la que las unidades son seleccionadas. Cuando el muestreo es probabilístico, las unidades de la muestra son seleccionadas de manera aleatoria mientras que en muestreo por cuotas se le deja al entrevistador elegir a quién muestrea. Esto resulta en sesgo de selección. Para hacer inferencias acerca de la población a partir de una muestra extraida por cuotas no probabilística es necesario suponer que las personas que fueron seleccionadas son similares a las que no lo fueron. Este supuesto es fuerte y raramente válido.

Pese a que, en general, muestras no probabilísticas no permiten que se realicen afirmaciones sobre toda la población de inferencia, tienen muchas ventajas que se enlistan a continuación:

- Son relativamente más baratas.
- Son más fáciles de administrar.
- Tienen la característica deseable de satisfacer proporciones de la población.
- Son más adecuadas para estudios cualitativos profundos cuyo objetivo es entender fenómenos sociales complejos, por ejemplo, véase Marshall (1996); Small (2009).
- La literatura indica que son apropiadas si se está interesado solamente en los casos específicos estudiados.
- La inferencia basada en diseño tiene la limitante de ser asintótica y tiene pocas herramientas para muestras pequeñas o áreas de estimación pequeñas. Para estos casos, el muestreo no probabilístico es una buena alternativa.
- Son una alternativa cuando no existe un marco muestral[4].

Algunas diferencias que es importante considerar entre muestreos probabilísticos y no probabilísticos son los siguientes:

- En muestreo probabilístico, cuando no hay un modelo para predecir los casos excluidos de la muestra (los errores antes mencionados 1. de cobertura y 3. por no respuesta), la verosimilitud calculada es básicamente no informativa. Esto los haría equivalentes en poder de inferencia a modelos no probabilísticos pero con el costo añadido del diseño y los costos de administración adicionales (Little, 2014, p. 422).
- El gran problema de muestreos no probabilísticos basados en cuotas deriva de no conocer los sesgos que pueden llevar a malas inferencias. Cuotas rutinarias de género, sexo y edad suelen funcionar bastante bien. Si bien, no se puede conocer si se está en la situación en la que se eliminan los sesgos o no, se puede decir lo mismo de las correcciones a la no respuesta en muestreos probabilísticos: a veces funcionan y a veces no. Cuando la no respuesta es muy alta[5], los grupos mal representados suelen ser minorías, personas de bajos ingresos o menor grado educativo.

[3] Es precisamente esta propiedad lo que permite realizar inferencias acerca de la población basadas en diseño: la aleatorización en la selección permite tener una estimación puntual y de intervalo para los parámetros que se desean conocer.

[4] Éste es el caso del levantamiento en Morelos pues no se buscaba realizar una encuesta en viviendas (para la cuál si existe un marco muestral) sino de personas en el espacio público. No existe un listado que permita realizar una selección aleatoria, es decir, no hay marco muestral.

[5] Generalmente, encuestas tradicionales que se realizan en colonias como las del Diagnóstico Participativo para el Programa Morelos: Territorio de Paz, con altos grados de inseguridad y que tratan temas delicados como percepción de seguridad, confianza en autoridades, entre otros; suele haber una alta tasa de no respuesta o de rechazo a la entrevista. Tal es el caso con la Encuesta de Línea de Base de la Evaluación de Impacto de Todos Por Acapulco realizada en 2012 por el Instituto Nacional de Salud Pública y que tuvo una tasa de respuesta del 63 % en el total de la muestra (Gutiérrez Reyes, Ruvalcaba, and Dyer, 2012).

En ninguno de los esquemas de muestreo pueden utilizarse estimadores directamente de la muestra para hablar de la población objetivo. En el caso de un muestreo probabilístico, se realizan inferencias basadas en diseño que requieren del cálculo de los factores de expansión con los que se pondera a cada unidad de análisis en la muestra[6]. Para muestreos no probabilísticos la única manera de realizar inferencias sobre la población es generando un modelo suficientemente robusto para controlar los sesgos de selección. La siguiente sección explicita los métodos más comunes.

3. Ajustes posteriores a los datos recolectados

Los ajustes posteriores a la realización de la encuesta son un elemento central para tener confianza en las estimaciones realizadas. Hay muchos procedimientos robustos para situaciones en los que el muestro se desvía de la probabilidad de selección o cuando hay problemas de cobertura o de no respuesta. Ponderar los datos es útil para ajustarlos a controles sociodemográficos.

Es importante notar que los procedimientos estadísticos estándar para hacer inferencia (cálculo de intervalos de confianza y pruebas de hipótesis) todavía requieren de una muestra probabilística. Cuando se realizan encuestas en la práctica, sobretodo en investigación de mercado y opinión pública, normalmente se violan los principios detrás del muestreo probabilístico ya que no se cuenta con un marco muestral, el tiempo o es demasiado caro diseñar un muestreo tradicional. Esto ha obligado a que los estadísticos especifiquen las condiciones en las que una muestra no probabilística funciona.

Los problemas para realizar inferencias a partir de muestras no probabilísticos se encuentran entre los más retadores en metodología de encuestas contemporánea (Vehovar and Manfreda, 2008, p. 184). Existen varios métodos que recientemente han sido desarrollados para realizar inferencia basada en modelos sobre diseños no probabilísticos. Entre éstos, se encuentran los siguientes tres: a) sample matching, b) máxima entropía y, c) regresión multinivel y postestratificación. El tercer método se especifica en la sección 4.

1. **Sample matching**: es un método para crear una muestra cuando se tiene un número grande de respuestas a una encuesta pero, posiblemente éstas no sean representativas. Cada muestra (individuos, hogares) puede ser "apareada" (en inglés, *matched*) a una unidad en un marco de muestreo de acuerdo a ciertas variables auxiliares[7]. La idea fundamental es que primero se selecciona una muestra objetivo del marco de muestreo utilizando algún tipo de muestreo probabilístico. Sin embargo, en vez de entrevistar a aquellos dentro de la muestra objetivo, se busca al más cercano del conjunto de encuestados disponibles para cada unidad dentro de la muestra objetivo. Colectivamente, las unidades "apareadas" son conocidas como la muestra de empate. El empate no necesita ser exacto y suele realizarse utilizando una función de distancia que mida la similitud entre pares de encuestados. Si la cantidad de encuestas disponibles es suficientemente grande, se garantiza que la muestra de empate tiene aproximadamente la misma distribución conjunta para las variables de empate que la muestra objetivo. Ejemplos del uso de esta técnica pueden encontrarse en Rivers (2007, p. 11) y en Hill, Lo, Vavreck, and Zaller (2007, p. 14).
2. **Máxima entropía**: el principio de máxima entropía se basa en utilizar la información disponible de la manera más eficiente. La entropía se utiliza para medir la incertidumbre que se tiene de la ocurrencia de una colección de eventos. Se elige entonces la distribución que maximiza la entropía sujeta a la muestra (por ejemplo, los momentos de la muestra) y, al mismo tiempo, la información no contenida en la muestra acerca de la variable aleatoria (Bernardini and Filippucci, 2000, p. 1687).

[6] La ventaja de un diseño es que para cada tipo existe una forma cerrada para realizar el cálculo del estimador puntual, el intervalo de confianza, el coeficiente de variación y demás medidas del error.

[7] Por ejemplo, en una encuesta a individuos en el que se cuenta con un marco muestral, se seleccionan aleatoriamente a elementos de éste según género, raza, edad e ingreso. Ahora bien, si uno de los individuos seleccionado aleatoriamente en el marco muestral cae en la categoría de *hombre, blanco, de 30 a 49 años y del último decil del ingreso*, entonces se busca en la muestra al elemento más cercano a estas características. Esto se hace para cada uno de los individuos seleccionados aleatoriamente del marco muestral.

4. Regresión multinivel y postestratificación

La **regresión multinivel y postestratificación** (MRP) permite desagregar los resultados de una encuesta sobre una serie de categorías y ayuda a corregir el sesgo de selección utilizando características demográficas y geográficas conocidas para hacer inferencias acerca de las personas en una región geográfica específica. Este método fue desarrollado por Gelman and Little (1997); Park, Gelman, and Bafumi (2006); Gelman and Hill (2007) y posteriormente retomado por Lax and Phillips (2009); Kastellec, Lax, and Phillips (2010). En este reporte se utiliza este método y se explica a mejor detalle a continuación.

El método MRP permite resolver dos tipos de problemas en encuestas: la estimación de áreas pequeñas[8] y/o los datos recolectados están afectados por el sesgo de selección. Estos dos problemas derivan en estimadores relativamente imprecisos de las categorías de interés. Estimadores de mayor precisión se pueden obtener utilizando una combinación apropiada de modelos de regresión multinivel y postestratificación (MRP).

4.1. El problema a solucionar

En general, el objetivo es determinar si, y hasta qué punto, una distribución para una variable de interés (Y) varía a través de las categorías de una variable (D)[9]. La distribución condicional de Y en cada categoría d de D puede escribirse como:

$$Y_d \sim f(\theta_d, \phi_d), d = 1, ..., J$$

donde

- $f(.)$ es una distribución de probabilidad genérica.
- θ_d es el valor esperado de la distribución.
- ϕ_d son parámetros suplementarios de la distribución de probabilidad (por ejemplo, la varianza).

Concentrémonos en el cálculo de θ_d. ¿En qué medida el valor esperado de Y varía en las J categorías de D? En regresión simple, el problema se reduce a estimar los J valores posibles de la función de regresión $E(Y|D = d) = \theta_d$. Esta colección de estimadores se denotará como $\theta \equiv \{\theta_d; d = 1, ..., J\}$. El problema es obtener un estimador preciso e insesgado para θ.

Se supone en adelante que:

- Las observaciones provienen de la población objetivo.
- Los datos de interés son recolectados sin error de medición.
- La única fuente de error de estimación es la varianza de la muestra[10] y la única posible fuente de estimación sistemática del error es el sesgo de selección (*selection bias*).

4.2. Estimación MRP

El estimador estándar de máxima verosimilitud de cada elemento θ_d de θ es

$$\hat{\theta}_d \equiv E(Y|\hat{D} = d) = \frac{\sum_{i=1}^{n_d} Y_i}{n_d}$$

[8] En inglés conocido como el *small area problem*, está presente cuando el número de observaciones válidas en una encuesta para una o más categorías es muy pequeña.

[9] Sin pérdida de generalidad, D puede representar una única variable categórica o una combinación de dos o más categóricas. Llamamos J al número de categorías en D y d a una categoría cualesquiera en D.

[10] En inglés, *sampling variance* e incluye la suma del error de cobertura, el error por no respuesta y el error de muestreo. Existe porque el valor de un estadístico varía entre las personas en la población objetivo y las encuestas miden únicamente a un subconjunto de ésta (Groves, 2004, p. 8-9).

donde n_d denota al número de observaciones en la muestra válidas en la categoría d de la variable D.

Cuando n_d es pequeño, $\hat{\theta_d}$ tiende a ser muy impreciso, es decir, genera estimadores muy variables de θ_d. La precisión de $\hat{\theta_d}$ decrece aun más si los datos tienen sesgo de selección, es decir, si las observaciones válidas son una muestra no aleatoria de la población objetivo y el proceso de selección en la muestra está asociado a una o más variables que también se asocian a Y. En estos dos casos, se puede obtener una buena estimación para θ utilizando MRP. Por la falta de corrección de sesgo en el estimador de máxima verosimilitud, no es apropiado para un muestreo no probabilístico basado en cuotas.

Se denotará al estimador por este método como $\tilde{\theta}$ y se obtiene con el siguiente proceso:

1. Identificación de una o más variables que pueden ser responsables del sesgo de selección. Sin pérdida de generalidad, la cuadrícula completa de clasificación generada por estas variables se trata como una única variable categórica G.
2. Se define un nuevo estimador $\gamma \equiv E(Y|D = d, G = g), d = 1, ..., J, g = 1, ..., G$[11].
3. Se utiliza un modelo de regresión multinivel apropiadamente especificado para estimar γ.
4. El paso de postestratificación utiliza el modelo generado en el paso 3. Se computa el estimador MRP para cada elemento θ_d de θ como la suma ponderada del subconjunto apropiado de $\hat{\gamma}$.

$$\tilde{\theta_d} = \sum_{g=1}^{G} \hat{\gamma_{d,g}} w_{g|d}$$

donde $w_{g|d} = \frac{N_{g,d}}{N_d}$. El numerador es el número de miembros de la población objetivo que pertenecen simultáneamente a la categoría g y d. El denominador es el número de miembros en la población objetivo que pertenecen a la categoría d.

Ventajas del método

- El uso de la regresión multinivel incrementa la precisión del estimador.
- Si G se define adecuadamente, la postestratificación ayuda a decrecer el error por sesgo de selección.
- MRP es un estimador relativamente preciso para θ.

Desventajas del método

- Se necesitan datos poblacionales para toda la clasificación DxG lo cuál limita la definición de G.
- Para obtener buenos estimadores de γ, el modelo de regresión multinivel debe ser especificado con mucho cuidado. Sin embargo, esta limitación aplica para cualquier modelo.

[11] D es la variable de interés, es decir, las respuestas a una pregunta específica de la encuesta. G son todas las posibles combinaciones de categorías de las variables elegidas en el paso 1. Por ejemplo, si las variables elegidas son género y ocupación (trabaja o no) entonces G tiene como categorías: mujer empleada, mujer desempleada, hombre empleado, hombre desempleado.

5. Ejemplo

Para ejemplificar los pasos, ventajas y limitaciones señalados en el apartado anterior, a continuación se realiza un ejemplo completo de estimación puntual para las encuestas levantadas en Morelos en 2013 y 2014.

5.1. Paso 1

La identificación de variables que pueden ser responsables del sesgo es lo más importante del análisis. Sin embargo, es el elemento más limitado pues los datos censales recolectados por INEGI no están disponibles a nivel individual. La información ya se encuentra agregada a nivel manzana y, por ende, se depende de las categorías que el Instituto generó para definir G. A nivel manzana es posible especificar únicamente 5 modelos, es decir, definir cinco cuadrículas G:

- Edad y unidad geográfica
- Condición de ocupación y unidad geográfica
- Escolaridad y unidad geográfica
- Género, edad y unidad geográfica
- Condición de ocupación, género y unidad geográfica

El levantamiento en Morelos y las cuotas por edad fueron realizadas a nivel colonia. Ésta no es una unidad administrativa controlada por INEGI. Se definieron polígonos geográficos para cada colonia, se crearon cuotas de edad para cada una y se identificaron puntos de afluencia en los cuales se realizó el levantamiento. Lo primero que se debe generar es una base de datos censal para las colonias en el estudio. Para esto, se identifican las manzanas que están contenidas en cada colonia y se suman los datos reportados para cada manzana[12]. Posteriormente, deben generarse las matrices de datos agregados según cada categoría G a modelar.

Se transforman los datos con el siguiente código en R.

```
library(plyr)
library(dplyr)
library(tidyr)
library(assertthat)
census <- readRDS("data/mrp_census.rds")
census <- census[!duplicated(census$idcolonia),]

################################################################################
# G: edad x colonia (dim=4x63)
mod.1 <- census %>%
    mutate("e12a17"=pob7+pob9
           , "e18a29"=pob11-pob9
           , "e30a49"=pob14
           , "e50mas"=pob15+pob23) %>%
    dplyr::select(idcolonia, e12a17, e18a29, e30a49, e50mas) %>%
    tidyr::gather(key, value, -idcolonia)

# Se eliminan valores nulos (no debe haberlos)
mod.1 <- na.omit(mod.1)
mod.1 <- dplyr::rename(mod.1, edad=key)
# Se verifica que las dimensiones de la cuadrícula sean apropiadas
```

[12] Es importante señalar que INEGI omite datos para manzanas en las cuales la población es tan pequeña que puede identificarse al individuo a partir de la información censal. Sin embargo, el estudio se concentró en áreas urbanas en las que esto no sucede a menudo.

```r
assert_that(dim(mod.1)[1]==63*4)

##############################################################################################
# G: trabajo x colonia
mod.2 <- census %>%
    mutate(trabaja=eco4
           , notrabaja=pob19-eco4) %>%
    dplyr::select(idcolonia, trabaja, notrabaja) %>%
    tidyr::gather(key, value, -idcolonia)

# Se eliminan valores nulos (no debe haberlos)
mod.2 <- na.omit(mod.2)
mod.2 <- dplyr::rename(mod.2, ocupacion=key)
# Se verifica que las dimensiones de la cuadrícula sean apropiadas
assert_that(dim(mod.2)[1]==63*2)

##############################################################################################
# G: escolaridad x colonia
mod.3 <- census %>%
    mutate(ninguno=edu31
           , primaria=edu34
           , secundaria = edu37
           , prepaomas = edu40
           ) %>%
    dplyr::select(idcolonia, ninguno, primaria, secundaria, prepaomas) %>%
    tidyr::gather(key, value, -idcolonia)

# Se eliminan valores nulos (no debe haberlos)
mod.3 <- na.omit(mod.3)
mod.3 <- dplyr::rename(mod.3, escolaridad=key)
# Se verifica que las dimensiones de la cuadrícula sean apropiadas
assert_that(dim(mod.3)[1]==63*4)

##############################################################################################
# G: genero x edad x colonia
mod.4 <- census %>%
    mutate(
        mujer_e12a17 = pob37 + pob39
        , mujer_e18a29 = pob41 - pob39
        , mujer_e30a49 = pob45
        , mujer_e50omas = pob46 + pob54
        , hombre_e12a17 = pob63 + pob65
        , hombre_e18a29 = pob67 - pob65
        , hombre_e30a49 = pob70
        , hombre_e50omas = pob71 + pob79
        ) %>%
    dplyr::select(idcolonia, mujer_e12a17, mujer_e18a29, mujer_e30a49,
                  mujer_e50omas, hombre_e12a17, hombre_e18a29, hombre_e30a49,
                  hombre_e50omas) %>%
    tidyr::gather(key, value, -idcolonia) %>%
    tidyr::separate(key, c("genero", "edad"), sep="_")

# Se eliminan valores nulos (no debe haberlos)
```

```r
mod.4 <- na.omit(mod.4)
# Se verifica que las dimensiones de la cuadrícula sean apropiadas
assert_that(dim(mod.4)[1]==63*4*2)

##############################################################################
# G: genero x ocupacion x colonia
mod.5 <- census %>%
    mutate(
        mujer_trabaja = eco5
        , mujer_notrabaja = pob50 - eco5
        , hombre_trabaja = eco6
        , hombre_notrabaja = pob75 - eco6
        ) %>%
    dplyr::select(idcolonia, mujer_trabaja, mujer_notrabaja, hombre_trabaja,
                  hombre_notrabaja) %>%
    tidyr::gather(key, value, -idcolonia) %>%
    tidyr::separate(key, c("genero", "ocupacion"), sep="_")

# Se eliminan valores nulos (no debe haberlos)
mod.5 <- na.omit(mod.5)
# Se verifica que las dimensiones de la cuadrícula sean apropiadas
assert_that(dim(mod.5)[1]==63 * 2 * 2)
```

El ejemplo completo utilizará el modelo 4, es decir, la cuadrícula por colonia, género y edad. Ésta incluye los datos de 63 colonias aquí llamada *idcolonia*. La variable de *edad* para el censo se puede categorizar en 4: de 12 a 17 años, 18 a 29, 30 a 49 y 50 o más. La variable de género se codifica como hombre o mujer.

Los datos de la encuesta correspondientes a las categorías que conforman G deben ser codificados y nombrados de la misma manera que las empleadas en los datos censales. El código en `R` es como sigue:

```r
##############################################################################
## Cargamos librerias
library(plyr)
library(dplyr)
library(tidyr)
library(stringr)
##############################################################################
## Lectura de datos
mor_13 <- readRDS("data/mor_13.rds")
mor_14 <- readRDS("data/mor_14.rds")
##############################################################################
## Funciones auxiliares para recategorizar las variables para postestratificacion
# Generamos edad identica
gedad <- function(x){
    ifelse(x > 130, NA,
           ifelse(x > 50, "e50mas",
                  ifelse(x > 30, "e30a49",
                         ifelse(x > 18, "e18a29",
                                ifelse(x > 12, "e12a17", NA)))))
}
# Generamos escolaridad identica
gescolaridad <- function(x){
    x <- as.character(x)
    ifelse(x == "Secundaria", "secundaria",
```

```
           ifelse(x == "Primaria", "primaria",
                  ifelse(x == "Licenciatura", "prepaomas",
                         ifelse(x == "Posgrado", "prepaomas", "ninguno"))))
}
# Generamos ocupacion identica
gocupacion13 <- function(x){
    x <- as.character(x)
    ifelse(x == "No", "notrabaja",
           ifelse(x == "Si", "trabaja", NA))
}
gocupacion14 <- function(x){
    x <- as.character(x)
    ifelse(x =="Estudias y trabajas", "trabaja",
           ifelse(x == "Trabajas", "trabaja", "notrabaja"))
}
# Generamos genero identico
ggenero <- function(x){
    str_trim(tolower(as.character(x)))
}
##############################################################################
## Transformacion de tablas
mor_13 <- mor_13 %>%
    mutate(edad = gedad(edad),
           escolaridad = gescolaridad(escolaridad),
           ocupacion = gocupacion13(trabajas),
           genero = ggenero(genero)
           )
mor_14 <- mor_14 %>%
    mutate(edad = gedad(edad),
           escolaridad = gescolaridad(escolaridad),
           ocupacion = gocupacion14(ocupacion),
           genero = ggenero(genero)
    )
##############################################################################
## Recodificacion de variables para mrp *las respuestas deben ser 0 o 1*
gdelito <- function(x){
    x <- as.character(x)
    ifelse(x == "Sí", 1, 0)
}

# Definición de la variable D, en este caso: En el último año, ¿usted ha sido
# víctima de algún delito?

# En morelos 13, pregunta 17
mor_13 <- mutate(mor_13,
                 victimizacion = gdelito(p17))
# En morelos 14, pregunta 27
mor_14 <- mutate(mor_14,
                 victimizacion = gdelito(p27))
```

En las últimas líneas de código, se categoriza la variable D elegida: *En el último año, ¿has sido víctima de algún delito?* Se recodifica como 1 para "Sí" y 0 para "No"[13]

[13]La implementación del método en R solo acepta variables dicotómicas para J.

5.2. Paso 2

Es una desventaja el no poder generar una mayor cantidad de cuadrículas para los datos censales. Lo ideal sería que, para cada variable D se pudieran elegir las variables más apropiadas para eliminar los sesgos en las estimaciones. Lax and Phillips (2013) ofrecen múltiples recomendaciones para una adecuada especificación del modelo. Debido a que se emplean modelos jerárquicos, es posible utilizar datos estatales o municipales para robustecer el modelo generado. Mas aún, es posible especificar el modelo de manera bayesiana lo que permite extender la potencia de inferencia sobretodo cuando los datos son muy escasos, este no es el caso de análisis[14].

Como se mencionó anteriormente, aquí se utiliza el modelo 4 con cuadrícula por género, edad y colonia.

5.3. Paso 3

En `R` hay una implementación de MRP, ejemplo del cuál se utiliza a continuación. Ésta admite mayor complejidad en la especificación del modelo jerárquico (Malecki, Lax, Gelman, and Wang, 2013; Malecki, Lax, and Gelman, 2013).

```
## Cargo datos: se utilizan los codigos que se mostraron en el paso 1
source("src/clean_census.r", chdir=T)
source("src/clean_encuesta.r", chdir=T)

## Cargo librerias
library(mrpdata)
library(mrp)

# Implementacion para 2013
mrp.simple13 <- mrp(victimizacion ~ idcolonia + edad + genero,
                    data=mor_13,
                    population=mod.4,
                    pop.weights="value")
# Implementacion para 2014
mrp.simple14 <- mrp(victimizacion ~ idcolonia + edad + genero,
                      data=mor_14,
                      population=mod.4,
                      pop.weights="value")
```

5.4. Paso 4

Para obtener los resultados de victimización después de la postestratificación, es posible realizar de forma muy sencilla los llamados:

```
options(digits=4)
# Estimaciones para género y edad 2013
genero.edad13 <- 100*poststratify(mrp.simple13, ~ genero+edad)
# Estimaciones para género y edad 2014
genero.edad14 <- 100*poststratify(mrp.simple14, ~ genero+edad)
```

[14] Para mayor detalle de la especificación de modelos multinivel ver Gelman and Hill (2007).

	e12a17	e18a29	e30a49	e50omas
hombre	18.4638	21.4390	27.861	22.353
mujer	17.1970	19.9539	26.118	20.921
Diferencia	-0.5953	0.5541	-4.409	-1.355
hombre	17.8685	21.9930	23.453	20.998
mujer	14.9715	18.4999	19.797	17.683
Diferencia	-2.2255	-1.4539	-6.322	-3.238

Cuadro 1: Comparación de estimaciones para 2013-2014 para género y edad en el área de estudio.

Para estimaciones de victimización a nivel colonia, simplemente se realiza lo siguiente:

```
options(digits=4)
# Estimaciones para colonia y género 2013
gec13 <- 100*poststratify(mrp.simple13, ~ idcolonia+genero)
# Estimaciones para colonia y género 2014
gec14 <- 100*poststratify(mrp.simple14, ~ idcolonia+genero)
```

Colonia	Hombres.13	Hombres.14	Dif.Hombres	Mujeres.13	Mujeres.14	Dif.Mujeres
Acatlipa	24.02	22.81	-1.2083	22.67	19.36	-3.3147
Lomas del Carril	17.60	20.87	3.2721	16.47	17.60	1.1336
Calera chica	23.84	19.05	-4.7892	22.31	15.96	-6.3529
Centro	25.08	23.16	-1.9219	23.46	19.50	-3.9586
Cuauhtemoc cardenas	23.49	18.21	-5.2794	22.00	15.11	-6.8916
El Porvenir	23.46	26.17	2.7056	22.17	22.25	0.0811
El campanario	23.53	21.01	-2.5182	22.47	17.90	-4.5723
Huizachera	26.36	19.56	-6.8064	24.36	16.20	-8.1576
La rosa	24.29	24.89	0.5994	22.92	21.16	-1.7602
Los tarianes	24.23	23.50	-0.7271	22.53	19.77	-2.7619
Miguel Hidalgo	28.39	28.72	0.3294	26.54	24.50	-2.0405
Morelos	27.85	27.02	-0.8248	26.27	23.04	-3.2348
Otilio Montaño	24.11	25.98	1.8661	22.48	21.99	-0.4990
Pedregal	23.82	23.43	-0.3926	21.65	19.50	-2.1504
Pinos Tejalpa	22.09	24.38	2.2939	20.27	20.32	0.0510
San Cristóbal los Lirios	23.75	20.89	-2.8606	24.12	19.05	-5.0736
San lucas	23.70	31.15	7.4434	22.23	26.75	4.5231
Tejalpa	30.59	25.67	-4.9238	28.24	21.46	-6.7862
Tejalpa	23.75	21.43	-2.3137	22.30	18.07	-4.2261
Vicente Guerrero	23.73	20.90	-2.8322	22.70	17.63	-5.0742

Colonia	Hombres.13	Hombres.14	Dif.Hombres	Mujeres.13	Mujeres.14	Dif.Mujeres
Vista Hermosa	23.88	21.10	-2.7848	22.67	17.86	-4.8070
Acapantzingo	23.77	18.84	-4.9299	22.48	15.81	-6.6706
Ahuatepec	19.14	20.63	1.4899	17.90	17.35	-0.5523
Altavista	22.88	19.51	-3.3732	21.40	16.24	-5.1603
Ampliación Chapultepec	23.36	20.84	-2.5147	21.74	17.39	-4.3515
Antonio Barona Centro	24.88	23.33	-1.5493	23.47	19.78	-3.6906
Atlacomulco	24.03	22.88	-1.1480	22.25	19.21	-3.0417
Carolina	27.82	20.40	-7.4246	26.51	17.20	-9.3137
Chamilpa	25.92	24.46	-1.4592	24.21	20.61	-3.5971
Emiliano Zapata	33.17	23.75	-9.4141	31.44	20.14	-11.3018
Gualupita	23.84	26.99	3.1501	22.22	22.82	0.5995
Lagunilla del salto	23.91	16.60	-7.3166	22.38	13.79	-8.5899
Las Granjas	23.71	22.20	-1.5135	22.51	18.89	-3.6148
Lienzo Charro	27.31	20.53	-6.7818	25.85	17.34	-8.5029
Margarita maza de juarez	24.26	23.47	-0.7854	22.46	19.78	-2.6798
Ocotepec	27.84	16.79	-11.0492	26.19	13.99	-12.2026
Patios de la estacion	23.35	17.68	-5.6700	22.17	14.92	-7.2565
San Antón	23.45	23.00	-0.4581	21.58	19.05	-2.5262
San cristobal	23.88	23.02	-0.8622	22.63	19.50	-3.1281
San miguel acapantzingo	23.94	21.74	-2.2000	22.34	18.23	-4.1099
Santa maria ahuacatitlan	23.86	21.54	-2.3188	22.34	18.09	-4.2494
Satélite	27.06	23.96	-3.1018	25.24	20.14	-5.1011
Tepepan	23.84	19.99	-3.8494	22.56	17.01	-5.5553
Vicente Estrada Cajigal	28.28	21.41	-6.8653	26.71	18.08	-8.6305
Villa santiago	23.97	23.08	-0.8958	22.34	19.38	-2.9626
prados de Cuernavaca	24.03	20.09	-3.9372	22.26	16.72	-5.5437
Año de Juárez	21.93	22.59	0.6584	20.76	19.17	-1.5956
Casasano	23.73	20.97	-2.7602	22.46	17.76	-4.7089
Centro	23.84	20.85	-2.9978	22.33	17.50	-4.8218
Cuauhtémoc	25.28	23.41	-1.8697	24.00	19.95	-4.0551
Cuautlixco	23.27	19.37	-3.8997	21.86	16.25	-5.6078
Guadalupe victoria	24.00	21.77	-2.2277	22.63	18.40	-4.2350
Hermenegildo Galeana	21.91	22.10	0.1935	20.49	18.63	-1.8656
Infonavit tetelcingo	25.04	23.15	-1.8945	23.25	19.51	-3.7396
Iztaccíhuatl	19.70	17.51	-2.1931	18.49	14.62	-3.8721
Juan Morales	19.24	18.61	-0.6292	17.17	15.07	-2.1066

Colonia	Hombres.13	Hombres.14	Dif.Hombres	Mujeres.13	Mujeres.14	Dif.Mujeres
Lázaro Cárdenas	23.89	21.02	-2.8755	22.45	17.68	-4.7716
Paraiso	23.69	16.91	-6.7860	22.35	14.18	-8.1683
Patria Libre	18.76	20.41	1.6457	17.57	17.18	-0.3940
Peña flores	23.56	16.11	-7.4445	22.45	13.56	-8.8915
Plan de Ayala	22.07	23.21	1.1381	20.63	19.60	-1.0296
Tetelcingo	19.48	16.08	-3.3971	18.28	13.47	-4.8076

Cuadro 2: Estimaciones para 2013-2014 por género y colonia.

6. Bibliografía

[1] R. Bernardini and C. Filippucci. "Inference from non-random samples: a maximum entropy approach". (2000). .

[2] R. D. Fricker. "Sampling methods for web and e-mail surveys". In: *N. Fielding* (2008), pp. 195-216.

[3] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Ed. by C. C. U. Press. Vol. 18. 2007.

[4] A. Gelman and T. C. Little. "Poststratification into many categories using hierarchical logistic regression". (1997).

[5] R. M. Groves. *Survey errors and survey costs*. Vol. 536. John Wiley & Sons, 2004.

[6] J. Gutiérrez Reyes, A. Ruvalcaba and D. Dyer. *Levantamiento, captura y análisis de línea base que servirá para medir el impacto de los proyectos de previsión social del delito con participación ciudadana derivados del SUBSEMUN 2012. Reporte de estadísticas descriptivas de línea base.* Tech. rep. Instituto Nacional de Salud Pública, ago. 2012.

[7] S. J. Hill, J. Lo, L. Vavreck and J. Zaller. "The opt-in Internet panel: Survey mode, sampling methodology and the implications for political research". In: *Unpublished manuscript at the University of California, Los Angeles, California* (2007). .

[8] J. P. Kastellec, J. R. Lax and J. Phillips. "Estimating state public opinion with multi-level regression and poststratification using R". In: *Unpublished manuscript, Princeton University* (2010). .

[9] J. R. Lax and J. H. Phillips. "Gay rights in the states: Public opinion and policy responsiveness". In: *American Political Science Review* 103.03 (2009), pp. 367-386.

[10] J. R. Lax and J. H. Phillips. "How should we estimate sub-national opinion using MRP? preliminary findings and recommendations". In: *Midwest Political Science Association* (2013).

[11] R. J. Little. "Survey sampling: Past controversies, current orthodoxy, future paradigms". In: *Past, Present, and Future of Statistical Science* (2014), p. 411.

[12] M. Malecki, J. R. Lax and A. Gelman. *mrpdata: Supplemental Data for Multilevel Regression and Poststratification*. R package version 1.0. 2013.

[13] M. Malecki, J. R. Lax, A. Gelman and W. Wang. *mrp: Multilevel Regression and Poststratification*. R package version 1.0-1. 2013.

[14] M. N. Marshall. "Sampling for qualitative research". In: *Family practice* 13.6 (1996), pp. 522-526.

[15] D. K. Park, A. Gelman and J. Bafumi. "State level opinions from national surveys: Poststratification using multilevel logistic regression". In: *Public opinion in state politics* (2006), pp. 209-28.

[16] D. Rivers. "Sampling for web surveys". In: *Joint Statistical Meetings*. 2007. .

[17] M. L. Small. "How many cases do I need?'On science and the logic of case selection in field-based research". In: *Ethnography* 10.1 (2009), pp. 5-38.

[18] V. Vehovar and K. L. Manfreda. "Overview: online surveys". In: *The SAGE handbook of online research methods* (2008), pp. 177-194.