Procesamiento de tablas de datos

Análisis de olas de calor y efectos en salud perinatal

José Daniel Conejeros jdconejeros@uc.cl

Tabla de datos de nacimiento

El total de nacimientos en Chile desde 1992 a 2020 es 7,489,864. La data fue alimentada con información adicional para los años 2019 y 2020.

Filtros de primer orden

- Filtro Nº1: se trabaja con los nacimientos solo para las comunas santiago. N=2,895,207 para 52 comunas en toda la Región Metropolitana.

```
births <- births %>% filter(comuna>=13000 & comuna<14000) # 2,724,086
# 89,770
births19 <- births19 %>% filter(comuna>=13000 & comuna<14000) %>% mutate(region=13)
# 81,351
births20 <- births20 %>% filter(comuna>=13000 & comuna<14000) %>% mutate(region=13)
```

Tenemos nacimientos desde el 1992-01-01 hasta el 2020-12-31 con 41 casos sin observaciones.

- Filtro Nº2: aplicamos `drop_na(date_nac)` para remover las observaciones donde no tenemos la fecha de nacimiento. Esto representa una pérdida de 41 observaciones, en otras palabras, el 0.001% de la muesta.
- Filtro Nº3: aplicamos `drop_na(weeks)` para remover las observaciones donde no tenemos la cantidad de semanas gestacionales. Sin esta información no podemos obtener las semana de exposición durante el proceso gestacional. Esto representa una pérdida de 2695 observaciones, en otras palabras, el 0.093% de la muesta original.
- Filtro Nº4: aplicamos `drop_na(weeks)` para remover las observaciones donde no tenemos el identificador de la comuna dentro de la Región Metropolitana. Sin esta información, no podemos cruzar con la tabla de datos de temperatura y olas de calor. Esto representa una pérdida de 0 observaciones. Todas tienen identificada la comuna. Hasta ahora tenemos un **N=2,892,471**.

A partir de esto construímos las variables que nos indican el inicio y el término del proceso gestacional:

```
births <- births %>%
  mutate(id=1:n()) %>%
  mutate(date_start = date_nac - weeks(semanas-1),
  date_end = date_nac)

# Primer caso de ejemplo
# weeks id  date_start  date_end
# 38    1   1992-06-27  1993-03-20
```

Filtros de segundo orden

Aplicamos un ajuste de las variables con missing values a partir del libro de códigos. En algunos casos estos missing values se codificaran como "Unknown" para no perder la información.

```
births <- births %>%
  mutate(
    size=if_else(talla==99, NA_real_, talla),
    age_mom=if_else(edad_madre==99, NA_real_, edad_madre),
    educ_mom=if_else(nivel_madre==9, NA_real_, nivel_madre),
    job_mom=if_else(activ_madre %in% c(9), NA_real_, activ_madre+1),
    age_dad=if_else(edad_padre==99, NA_real_, edad_padre),
    educ_dad=if_else(nivel_padre==9, NA_real_, nivel_padre),
    job_dad=if_else(activ_padre %in% c(3,9), NA_real_, activ_padre+1)
    )
```

Tomando las variables generadas se realiza un proceso de fltrado de segundo orden:

- Filtro Nº5: se excluyen gestantes con menos de 12 años de edad y mayores de 50 años de edad. `filter(age_mom>=12 & age_mom<=50)` esto remueve 641 observaciones. Un 0.022% menos de la data restante para la RM.
- Filtro Nº6: se excluyen gestantes con menos 28 semanas de gestación. `filter(weeks >= 28)` esto remueve 11641 observaciones. Un 0.403% menos de la data restante para la RM.
- Filtro Nº7: se excluyen gestantes con más de 1 nacimiento. `filter(tipo_parto==1)` esto remueve 56940 observaciones. Un 2.017% menos de la data restante para la RM.

Las observaciones que quedan después de este proceso son **N=2,823,249**.

Construcción de covariables para la data de nacimientos

En esta del análisis se codificaron valores 9 y missing como "Unknown" para no perder esa información e incorporar el patrón de no respuesta dentro del modelamiento. A continuación se presenta la construcción de las covariables:

```
# El flujo solo considera la construcción de variables
births <- births %>%
  mutate(sex=factor(sexo, levels=c(1,2,9), labels=c("Boy", "Girl", "Unknown"))) %>%
  # Edad, educación y ocupación del padre es equivalente al de la madre.
  mutate(
    age_group_mom=case_when(
      age_mom <= 20 ~ 1,
      age_mom > 20 & age_mom <= 29 ~ 2,
      age_mom >= 30 & age_mom <= 39 ~ 3,
      age_mom >= 40 & age_mom <= 49 ~ 4,
      age_mom >= 50 ~ 5,
      TRUE ~ 6
    ),
    age_group_mom=factor(age_group_mom,
                         levels=c(1:6),
                         labels=c("<=20", "20-29", "30-39", "40-49", ">=50", "Unknown")),
    educ_group_mom = case_when(
      educ_mom == 1 ~ 4, # College
      educ_mom == 2 ~ 3, # Secondary
      educ_mom == 3 ~ 3, # Secondary
      educ_mom == 4 ~ 2, # Primary
      educ_mom == 5 ~ 1, # No educaction
      TRUE ~ 5, #Unknow
    ),
    educ_group_mom = factor(educ_group_mom,
                            levels = c(1:5),
                            labels = c("Non education", "Primary", "Secondary",
                             "College", "Unknown")),
    job_group_mom = if_else(is.na(job_mom), 4, job_mom),
    job_group_mom = factor(job_group_mom,
                            levels = c(1,2,3,4),
                            labels=c("Not working", "Employed", "Unemployed", "Unknown"))
```

Filtros de tercer orden

Los filtros de tercer orden ajustan por valores ilógicos entre semanas de gestación y peso del bebe nacido. Se pueden considerar dos criterios:

```r
# Criterio USA
births <- births %>%
  filter(
    case_when(
      weeks == 28 ~ tbw >= 250 & tbw <= 2500,
      weeks == 29 ~ tbw >= 250 & tbw <= 2750,
      weeks == 30 ~ tbw >= 375 & tbw <= 3000,
      weeks == 31 ~ tbw >= 375 & tbw <= 3250,
      weeks == 32 ~ tbw >= 500 & tbw <= 3500,
      weeks == 33 ~ tbw >= 500 & tbw <= 3750,
      weeks == 34 ~ tbw >= 500 & tbw <= 4000,
      weeks == 35 ~ tbw >= 750 & tbw <= 4500,
      weeks == 36 ~ tbw >= 750 & tbw <= 5000,
      weeks == 37 ~ tbw >= 750 & tbw <= 5500,
      weeks >= 38 ~ tbw >= 1000 & tbw <= 6000,
      TRUE ~ FALSE
    )
  )
```

Este criterio remueve 749 observaciones que están fuera de rango para cada semana gestacional, más 165 observaciones sin información para el peso. Esto da una pérdida de 914 observaciones que corresponden al 2.017%.

Se considero una segunda definición a partir de percentiles de peso para año-semana de gestación. Esto remueve 26441 que corresponden a 0.937%, Por lo tanto se optó por la primera definición.

```r
births <- births %>%
    group_by(year_week1, weeks) %>%
  mutate(
    P2.5 = quantile(tbw, probs = 0.005, na.rm = TRUE),
    P97.5 = quantile(tbw, probs = 0.995, na.rm = TRUE),
    test = tbw >= P2.5 & tbw <= P97.5
  ) %>%
  ungroup() %>%
  filter(!test)
```

Outcomes de análisis

La principal variable de interés corresponde a los partos preterminos definidos como los partos con menos de 37 semanas. A partir de esto se construyen diferentes definiciones de partos preterminos.

```r
births <- births %>%
  mutate(birth_preterm = if_else(weeks < 37, 1, 0)) %>%
  mutate(birth_very_preterm = if_else(weeks >= 28 & weeks <32, 1, 0)) %>%
  mutate(birth_moderately_preterm = if_else(weeks >= 32 & weeks <33, 1, 0)) %>%
  mutate(birth_late_preterm = if_else(weeks >= 34 & weeks <37, 1, 0)) %>%
  mutate(birth_term = if_else(weeks >= 37 & weeks <42, 1, 0)) %>%
  mutate(birth_posterm = if_else(weeks >= 42, 1, 0))
```

El total de observaciones es **N=2,823,249**. En este caso cada fila corresponde a un gestante. Sin embargo, necesitamos expandir los datos a semanas de gestación para dimensionar la exposición.

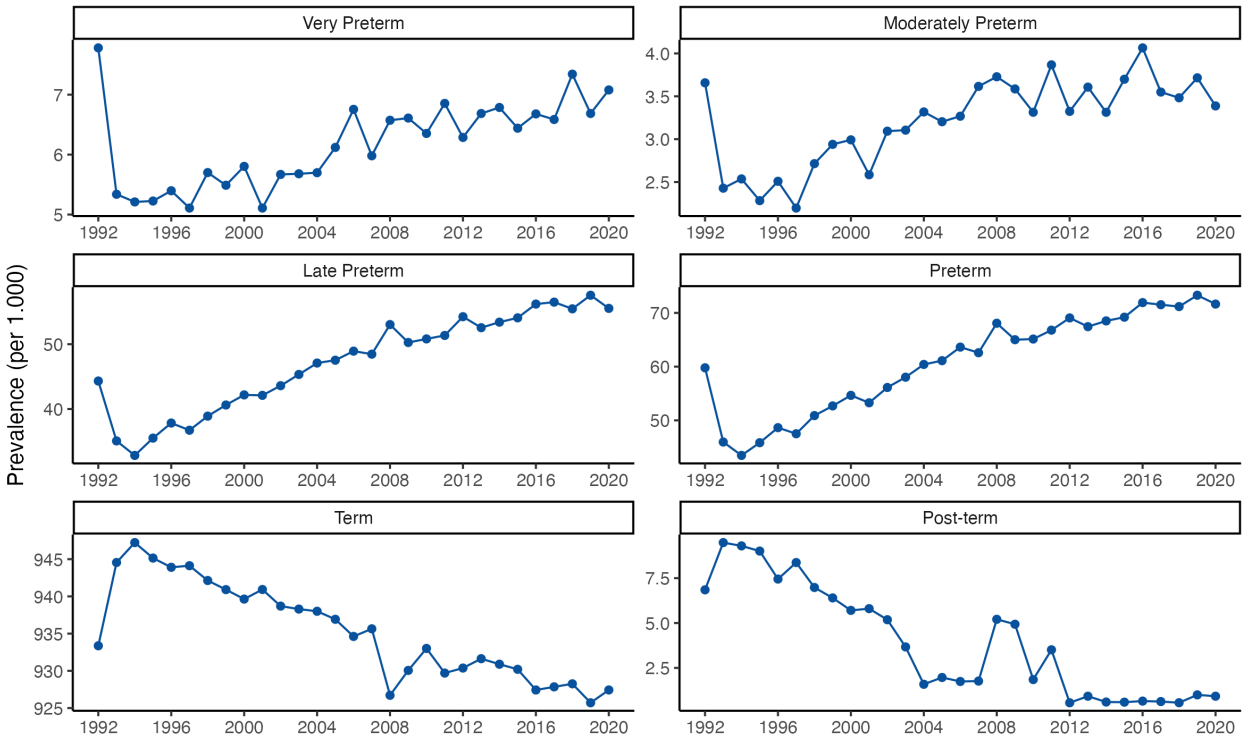
Tabla de datos de exposición

Fixed cohort bias

Los datos de exposión tienen su primera semana de gestación desde el "1991-03-13" al "2020-06-10". En primer instancia se procedió a aplicar el siguiente filtro sobre los datos:

```r
births <- births %>% filter(year_week1>=1992)
```

Se pierden 77,980 observaciones con fechas de inicio anteriores al año 1992.



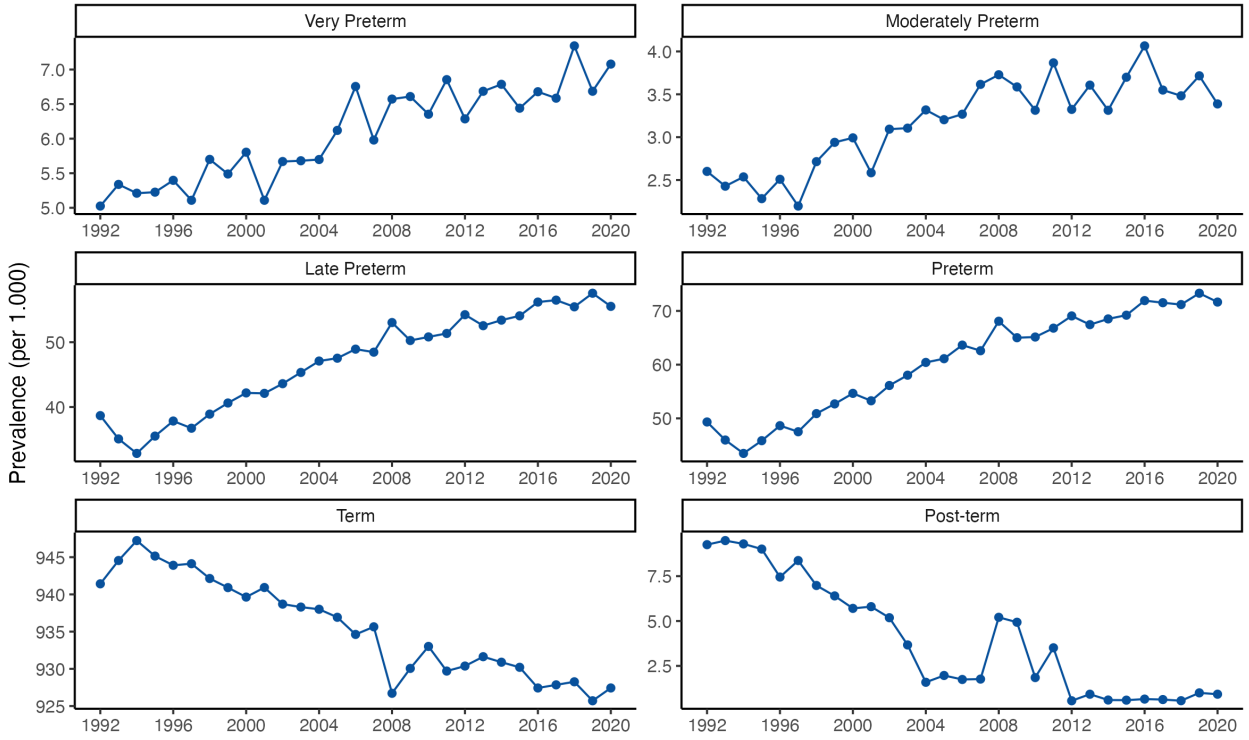
Aplicar un filtro sobre los datos de 1992 introduce un sesgo pues removemos artificialmente esas filas.

```
weeks <- rio::import("Descriptives/Start_ends_gestational_weeks.xlsx")
head(weeks[,c(1:4)], n=4) # Primeros nacimientos
```

```
  date_start_week_gest date_ends_week_gest weeks n_gestantes
1           1991-03-20          1992-01-01    42           4
2           1991-03-21          1992-01-02    42           4
3           1991-03-22          1992-01-03    42           7
4           1991-03-23          1992-01-04    42           4
```

```
tail(weeks[,c(1:4)], n=4) # Últimos nacimientos
```

```
       date_start_week_gest date_ends_week_gest weeks n_gestantes
115508           2020-06-08          2020-12-28    30           4
115509           2020-06-10          2020-12-23    29           2
115510           2020-06-15          2020-12-28    29           2
115511           2020-06-17          2020-12-23    28           2
```



Si los datos de temperatura tienen información desde "1991-01-01" no es necesario remover estas observaciones.

Expansión de las semanas gestacionales

A continuación se expande la tabla de datos para cada semana de gestación. Esto es necesario para realizar la imputación posterior con la información de olas de calor al final del proceso gestacional.

Este es un proceso largo que tarda alrededor de 50 minutos y expande la tabla de datos a un N= observaciones pues la unidad de análisis pasa a ser la semana gestacional.

```
births <- births %>%
  rowwise() %>%
  mutate(week_gest = list(seq.Date(date_start_week_gest, date_ends_week_gest, by = "week"))) %>
  unnest(week_gest) %>%
  group_by(id) %>%
  mutate(week_gest_num = paste0(abs(weeks - row_number())),
         week_gest_num = (weeks) - as.numeric(week_gest_num),
         date_start_week = (week_gest - (7 * abs(week_gest_num - row_number()))) - weeks(1), #(
         date_start_week = date_start_week + 1,
         date_end_week = week_gest - (7 * abs(week_gest_num - row_number()))
         ) %>% # ,(abs(week_gest_num - row_number())
  group_by(id) %>%
  distinct(week_gest_num, .keep_all = TRUE) %>%
  arrange(id, week_gest_num) %>%
  ungroup()
```

Con esto nos quedamos solo con las gestantes donde su últimas 4 semanas gestacionales estén en los meses de Noviembre, Diciembre, Enero, Febrero y Marzo:

```
dates_range <- function(date) {
  any(date >= as.Date(paste0(year(date)-1, "-11-01")) &
    fecha <= as.Date(paste0(year(date), "-03-31")))
}
## Last month -> Nov, Dic, Ene, Feb, Mar
births_last_month <- births %>%
  group_by(id) %>%
  filter(week_gest_num > (max(week_gest_num) - 4)) %>%
  filter(dates_range(date_end_week))

## Last week  -> Nov, Dic, Ene, Feb, Mar
births_last_week <- births %>%
  group_by(id) %>%
  filter(week_gest_num == max(week_gest_num)) %>%
  filter(dates_range(date_end_week))
```