Suplementary Material 1: Code to Reproduce Analysis and Plots

Richard Meitern

2023-03-22

Introduction

The aim of this document is to display the R code needed the reproduce the findings of the main text. This documents needs the twinR package to be installed to run.

```
#cleanup memory
gcstuff <- gc(verbose=FALSE); rm(gcstuff);
```

```
#get last birth adding function
source("./R/last_birth.R")

#simplified twinR summary tables
source("./R/twinR_summary.R")

#fix twinR compute predictions to do prediction with no lambda  as well
source("./R/twinR_predictions.R")

#simple convenience functions
source("./R/utils.R")
```

```
## Identify number of CPU cores available for parallel computing,
## note: using a large number may lead RAM to max out, so you may have to adjust
## that according to your infrastructure:
nb_cores <- min(c(50L, parallel::detectCores() - 1))

## Set option in spaMM:
spaMM::spaMM.options(nb_cores = nb_cores)
```

```
## Registered S3 methods overwritten by 'registry':
##   method               from
##   print.registry_field proxy
##   print.registry_entry proxy
```

Data import

The Estonian dataset has been formatted to include the same columns as *the data_births_all* dataset from the **twinR** package. The only difference is that the columns *pop* and *monthly* are excluded as these are constant.

```
#Import and preproccess Estonian Data

data_births_monthly_EE <- readRDS("./data/data_births_all_EE.rds")

#the twinR package expects population to be present
data_births_monthly_EE$pop <- "Estonia"

## Expand the birth level data for the fit of statistical models:
data_births_monthly_EE <- twinR::expand_data(data_births_monthly_EE)

data_births_monthly_EE <- add_last_birth(data_births_monthly_EE)

data_births_monthly_EE_not_last <- data_births_monthly_EE[!data_births_monthly_EE$last,]

#make the aggregates
dmm_EE <- twinR::aggregate_data(data_births_monthly_EE)
dmm_EE$prob_twin <- dmm_EE$twin_total / dmm_EE$births_total

dmm_EE_nl <- twinR::aggregate_data(data_births_monthly_EE_not_last)
dmm_EE_nl$prob_twin <- dmm_EE_nl$twin_total / dmm_EE_nl$births_total
```

```
ee_tbl <- twinR::build_data_summary.table(data_births_monthly_EE)
ee_tbl_nl <- twinR::build_data_summary.table(data_births_monthly_EE_not_last)
```

```
  eetbls <- rbind(ee_tbl[-1,4:ncol(ee_tbl)-1],ee_tbl_nl[-1,4:ncol(ee_tbl_nl)-1])
eetbls$Subset <- c("All data", "No last birth")
knitr::kable(eetbls,
             caption = paste0("Data summary table of Estonian data",
                              " with and without last born child"))
```

Table 1: Data summary table of Estonian data with and without last born child

Maternal birth period	Mothers	Non-twinners	Twinners	Twinner rate (‰)	Offspring birth period	Births	Singleton births	Twin births	Twinning rate (‰)	Total births (min-median-max)	Subset
1850-1899	125575	119511	6064	48	1868-1948	417418	411026	6392	15	1-4-16	All data
1850-1899	98183	94386	3797	39	1868-1943	291843	287874	3969	14	1-4-15	No last birth

Data import Original

```
##Import and pre-proccess twinR package data

## Filter the raw data to only keep data with monthly resolution:
data_births_monthly <- twinR::filter_data(twinR::data_births_all)

## Expand the birth level data for the fit of statistical models:
data_births_monthly <- twinR::expand_data(data_births_monthly)

data_births_monthly <- add_last_birth(data_births_monthly)
data_births_monthly_not_last <- data_births_monthly[!data_births_monthly$last,]

dmm_orig <- twinR::aggregate_data(data_births_monthly)
dmm_orig$prob_twin <- dmm_orig$twin_total /dmm_orig$births_total

dmm_orig_nl <- twinR::aggregate_data(data_births_monthly_not_last)
```

```
dmm_orig_nl$prob_twin <- dmm_orig_nl$twin_total /dmm_orig_nl$births_total
```

```
knitr::kable(twinR::build_data_summary.table(data_births_monthly_not_last)[-2],
             caption = "Data sumary table without last born child")
```

Table 2: Data sumary table without last born child

Population	Maternal birth period	Mothers	Non-twinners	Twinners	Twinner rate (‰)	Offspring birth period	Births	Singleton births	Twin births	Twinning rate (‰)	Total births (min-median-max)	References
Finland East	1742-1899	778	711	67	86	1771-1938	3573	3503	70	19.6	1-6-16	Pettay et al. 2016; Pettay et al. 2018
Finland Lapland	1702-1884	669	619	50	75	1725-1915	2791	2739	52	18.6	1-5-12	Helle 2019
Finland SW-Archipelago	1709-1899	2320	2114	206	89	1732-1941	9399	9173	226	24.1	1-5-14	Haukioja et al. 1989; Lummaa et al. 1998
Finland West	1701-1899	4944	4604	340	69	1721-1939	25064	24693	371	14.8	1-7-15	Pettay et al. 2016; Pettay et al. 2018
Krummhörn	1705-1823	3364	3164	200	59	1725-1864	13895	13683	212	15.3	1-5-16	Gabler and Voland 1994
Sami Lapland	1703-1880	828	776	52	63	1729-1917	3901	3844	57	14.6	1-6-12	Helle et al. 2004; Helle 2019
Sweden Lapland	1721-1878	1715	1601	114	66	1749-1900	9163	9038	125	13.6	1-7-16	Sköld and Axelsson 2008; Sköld et al. 2011; Helle 2019
Switzerland	1700-1899	3902	3755	147	38	1720-1939	16757	16603	154	9.2	1-6-17	Evans et al. 2018
All the above	1700-1899	18520	17344	1176	64	1720-1941	84543	83276	1267	15.0	1-6-17	This paper

Fitting models

```
#' Fit Predictions
#'
#' This function fits a model using the given formula and dataset and
#' computes predictions. The model is fit using the \code{\link[spaMM]{fitme}}
#' function from the \strong{spaMM} package.
#'
#' @param dataset A data frame containing the data to be used for fitting the model.
#' @param formula A formula specifying the model to be fit.
#' @param predict Logical value indicating whether to do predictions. Default is TRUE.
#' @param nb_boot Number of bootstrap samples to use when computing predictions.
#'  Default is 1000.
#' @param predictionsDir Directory where precomputed predictions are stored.
#'  Default is "./data/predictions".
#'
#' @return A list containing the fitted model object and a data frame with
#'  computed predictions.
fitPredictions <- function(dataset, formula,
                           predict=TRUE,
                           nb_boot=1000,
                           predictionsDir = "./data/predictions"){

  if(!dir.exists(predictionsDir)) dir.create(predictionsDir)

  args <- list(formula = stats::as.formula(formula),
               data = dataset,
               family = stats::binomial(link = "logit"),
               method = "PQL/L")

  fit <-  twinR::fit_model_safely(timeout = Inf, .args = args)
```

```
  fitName <- deparse(substitute(dataset))
  #TODO maybe add formula also to the fitName like:
  #fitName <- paste0(form2str(stats::as.formula(formula)),fitName)

  predDataFname <- paste0(predictionsDir,"/",fitName ,"data_fig.rds")
  if(!file.exists(predDataFname) & predict){
    min_births <- min(dataset$births_total)
    max_births <- max(dataset$births_total)
    nd <- data.frame(births_total = min_births:max_births)
    data_fig <- compute_predictions(fit,
                                    newdata = nd,
                                    nb_boot = nb_boot)


    saveRDS(data_fig, predDataFname)

  } else {
    if(file.exists(predDataFname)){
      warning("Pre-computed predictions returned from file:\n", predDataFname,
            "\n If you want to re-run this time intensive step delete the file!")
    data_fig <- readRDS(predDataFname)
    } else {
      data_fig <- list(results = NULL)
    }

  }

  #garbage collecion after spaMM multi-core proccess
  gcstuff <- gc(verbose=FALSE)

  return(list(fit=fit, results=data_fig$results))
}
```

Full Data

```
## Estonia - mother level data
formula <- "cbind(twin_total, singleton_total) ~ 1 + births_total"
dmm_EE_fit <- fitPredictions(dmm_EE, formula)
```

```
## Warning in fitPredictions(dmm_EE, formula): Pre-computed predictions returned from file:
## ./data/predictions/dmm_EEdata_fig.rds
##  If you want to re-run this time intensive step delete the file!
```

```
knitr::kable(build_fit_summary.table(dmm_EE_fit$fit))
```

Type	Variable	Value	Cond. SE	t-value
fixed effects	(Intercept)	-4.2	0.03	-154.49
	births_total	0.0	0.01	-0.36
response family	binomial with logit link			
fit info	number of model parameters	2.0		

Type	Variable	Value	Cond. SE	t-value
	marginal log Likelihood	-24446.1		
	marginal AIC	48896.3		
	conditional AIC (cAIC)			
data info	number of fitted observations (N)	125575.0		

```
## Estonia - birth level data
formula <- "twin ~ 1 + poly(cbind(age, parity), 3) + (1|maternal_id)"
dbm_EE_fit <- fitPredictions(data_births_monthly_EE, formula, predict = F)
```

```
## If the 'ROI.plugin.glpk' package were installed,
## spaMM could properly check (quasi-)separation in binary regression problem.
## See help('external-libraries') if you have troubles installing 'ROI.plugin.glpk'.

## Increase spaMM.options(separation_max=<.>) to at least 4175 if you want to check separation (see 'he
```

```
knitr::kable(build_fit_summary.table(dbm_EE_fit$fit))
```

```
## [one-time computation of covariance matrix, which may be slow]
```

Type	Variable	Value	Cond. SE	t-value
fixed effects	(Intercept)	-4.23	0.03	-146.78
	poly(cbind(age, parity), 3)1.0	124.69	23.34	5.34
	poly(cbind(age, parity), 3)2.0	-115.54	14.20	-8.14
	poly(cbind(age, parity), 3)3.0	-67.80	14.91	-4.55
	poly(cbind(age, parity), 3)0.1	44.06	35.86	1.23
	poly(cbind(age, parity), 3)1.1	6070.97	20377.08	0.30
	poly(cbind(age, parity), 3)2.1	13475.80	14544.20	0.93
	poly(cbind(age, parity), 3)0.2	-6.92	26.95	-0.26
	poly(cbind(age, parity), 3)1.2	-11971.11	14297.44	-0.84
	poly(cbind(age, parity), 3)0.3	-12.04	12.98	-0.93
random effects	variance between name	0.52		
response family	binomial with logit link			
fit info	number of model parameters	11.00		
	marginal log Likelihood	-32714.88		
	marginal AIC	65451.75		
	conditional AIC (cAIC)			
data info	number of fitted observations (N)	417418.00		

```
## TwinR - mother level data
formula <- "cbind(twin_total, singleton_total) ~ 1 + births_total + (1|pop)"
dmm_orig_fit <- fitPredictions(dmm_orig, formula)
```

```
## Warning in fitPredictions(dmm_orig, formula): Pre-computed predictions returned from file:
## ./data/predictions/dmm_origdata_fig.rds
##  If you want to re-run this time intensive step delete the file!
```

```
knitr::kable(build_fit_summary.table(dmm_orig_fit$fit))
```

Type	Variable	Value	Cond. SE	t-value
fixed effects	(Intercept)	-3.83	0.10	-36.7
	births_total	-0.03	0.01	-3.9
random effects	variance between name	0.06		
response family	binomial with logit link			
fit info	number of model parameters	3.00		
	marginal log Likelihood	-5993.12		
	marginal AIC	11992.24		
	conditional AIC (cAIC)			
data info	number of fitted observations (N)	21290.00		

```
## TwinR - birth level data
formula <- "twin ~ 1 + poly(cbind(age, parity), 3) + (1|maternal_id) + (1|pop)"
dbm_orig_fit <- fitPredictions(data_births_monthly, formula, predict = F)
```

```
## Increase spaMM.options(separation_max=<.>) to at least 1059 if you want to check separation (see 'hel
```

```
knitr::kable(build_fit_summary.table(dbm_orig_fit$fit))
```

```
## [one-time computation of covariance matrix, which may be slow]
```

Type	Variable	Value	Cond. SE	t-value
fixed effects	(Intercept)	-4.10	0.11	-36.38
	poly(cbind(age, parity), 3)1.0	73.70	31.07	2.37
	poly(cbind(age, parity), 3)2.0	-61.18	17.22	-3.55
	poly(cbind(age, parity), 3)3.0	-47.11	16.42	-2.87
	poly(cbind(age, parity), 3)0.1	-0.94	40.14	-0.02
	poly(cbind(age, parity), 3)1.1	-4005.74	11889.93	-0.34
	poly(cbind(age, parity), 3)2.1	7709.69	8669.07	0.89
	poly(cbind(age, parity), 3)0.2	18.98	27.35	0.69
	poly(cbind(age, parity), 3)1.2	-4380.22	8120.86	-0.54
	poly(cbind(age, parity), 3)0.3	-18.63	14.09	-1.32
random effects	variance between name	0.48		
	variance between name	0.06		
response family	binomial with logit link			
fit info	number of model parameters	12.00		
	marginal log Likelihood	-8828.38		
	marginal AIC	17680.76		
	conditional AIC (cAIC)			
data info	number of fitted observations (N)	105833.00		

No Last Births Data

```
## Estonia
formula <- "cbind(twin_total, singleton_total) ~ 1 + births_total"
dmm_EE_nl_fit <- fitPredictions(dmm_EE_nl, formula)
```

```
## Warning in fitPredictions(dmm_EE_nl, formula): Pre-computed predictions returned from file:
## ./data/predictions/dmm_EE_nldata_fig.rds
##  If you want to re-run this time intensive step delete the file!
```

```
knitr::kable(build_fit_summary.table(dmm_EE_nl_fit$fit))
```

Type	Variable	Value	Cond. SE	t-value
fixed effects	(Intercept)	-4.36	0.03	-131.0
	births_total	0.02	0.01	2.7
response family	binomial with logit link			
fit info	number of model parameters	2.00		
	marginal log Likelihood	-15892.99		
	marginal AIC	31789.98		
	conditional AIC (cAIC)			
data info	number of fitted observations (N)	98183.00		

```
## Estonia - birth level data
formula <- "twin ~ 1 + poly(cbind(age, parity), 3) + (1|maternal_id)"
dbm_EE_nl_fit <- fitPredictions(data_births_monthly_EE_not_last, formula,
                                predict = F)
```

```
## Increase spaMM.options(separation_max=<.>) to at least 2919 if you want to check separation (see 'hel
```

```
knitr::kable(build_fit_summary.table(dbm_EE_nl_fit$fit))
```

```
## [one-time computation of covariance matrix, which may be slow]
```

Type	Variable	Value	Cond. SE	t-value
fixed effects	(Intercept)	-4.31	0.04	-118.27
	poly(cbind(age, parity), 3)1.0	118.86	22.47	5.29
	poly(cbind(age, parity), 3)2.0	-33.64	14.46	-2.33
	poly(cbind(age, parity), 3)3.0	-21.94	13.98	-1.57
	poly(cbind(age, parity), 3)0.1	54.44	37.75	1.44
	poly(cbind(age, parity), 3)1.1	-11865.19	17925.73	-0.66
	poly(cbind(age, parity), 3)2.1	9736.49	11627.66	0.84
	poly(cbind(age, parity), 3)0.2	19.59	29.17	0.67
	poly(cbind(age, parity), 3)1.2	-11242.39	11738.46	-0.96
	poly(cbind(age, parity), 3)0.3	-3.45	12.90	-0.27
random effects	variance between name	0.49		
response family	binomial with logit link			
fit info	number of model parameters	11.00		
	marginal log Likelihood	-20819.80		
	marginal AIC	41661.59		
	conditional AIC (cAIC)			
data info	number of fitted observations (N)	291843.00		

```
## TwinR
formula <- "cbind(twin_total, singleton_total) ~ 1 + births_total + (1|pop)"
dmm_orig_nl_fit <- fitPredictions(dmm_orig_nl, formula)
```

```
## Warning in fitPredictions(dmm_orig_nl, formula): Pre-computed predictions returned from file:
## ./data/predictions/dmm_orig_nldata_fig.rds
##  If you want to re-run this time intensive step delete the file!
```

```
knitr::kable(build_fit_summary.table(dmm_orig_nl_fit$fit))
```

Type	Variable	Value	Cond. SE	t-value
fixed effects	(Intercept)	-4.14	0.12	-34.85
	births_total	0.00	0.01	-0.09
random effects	variance between name	0.07		
response family	binomial with logit link			
fit info	number of model parameters	3.00		
	marginal log Likelihood	-4490.84		
	marginal AIC	8987.68		
	conditional AIC (cAIC)			
data info	number of fitted observations (N)	18520.00		

```
## TwinR - birth level data
formula <- "twin ~ 1 + poly(cbind(age, parity), 3) + (1|maternal_id) + (1|pop)"
dbm_orig_nl_fit <- fitPredictions(data_births_monthly_not_last, formula, predict = F)
```

```
## Increase spaMM.options(separation_max=<.>) to at least 846 if you want to check separation (see 'hel
```

```
knitr::kable(build_fit_summary.table(dbm_orig_nl_fit$fit))
```

```
## [one-time computation of covariance matrix, which may be slow]
```

Type	Variable	Value	Cond. SE	t-value
fixed effects	(Intercept)	-4.24	0.13	-33.43
	poly(cbind(age, parity), 3)1.0	85.90	31.19	2.75
	poly(cbind(age, parity), 3)2.0	-34.85	17.95	-1.94
	poly(cbind(age, parity), 3)3.0	-22.85	15.92	-1.43
	poly(cbind(age, parity), 3)0.1	-23.14	42.20	-0.55
	poly(cbind(age, parity), 3)1.1	4849.99	11442.24	0.42
	poly(cbind(age, parity), 3)2.1	2439.65	7666.86	0.32
	poly(cbind(age, parity), 3)0.2	7.32	29.51	0.25
	poly(cbind(age, parity), 3)1.2	1548.37	7323.01	0.21
	poly(cbind(age, parity), 3)0.3	-11.86	13.04	-0.91
random effects	variance between name	0.54		
	variance between name	0.07		
response family	binomial with logit link			
fit info	number of model parameters	12.00		
	marginal log Likelihood	-6480.49		
	marginal AIC	12984.99		

Type	Variable	Value	Cond. SE	t-value
	conditional AIC (cAIC)			
data info	number of fitted observations (N)	84543.00		

Plots

```
library(ggplot2)
#some nice colors
bc <- c("purple", "black", "navy", "darkgoldenrod2", "springgreen3", "gray")

#use new base theme that displays also grid lines
source("./R/twinR_theme.R")
```

Fig 1: Estonian vs TwinR Full Data

```
fig2_EE_plot_data <- dmm_EE_fit$results
fig2_orig_plot_data <- dmm_orig_fit$results
```

```
fig2_ext_orig <- ggplot() +
  geom_line(data=fig2_EE_plot_data,
              aes(y = estimate,  x=births_total, color="EE all"),  size = 1) +
    stat_summary(data=dmm_EE[dmm_EE$births_total <17, ],
                 aes(x=births_total, y=prob_twin, color="EE all", fill = "EE all"),
                 alpha=0.5,
                 fun.data=mean_se) +
    geom_ribbon(data=fig2_EE_plot_data,
                aes(y = estimate,  x=births_total, ymin = lwr, ymax = upr,
                    color="EE all", fill = "EE all"),
                alpha = 0.3) +
    geom_line(data=fig2_orig_plot_data,
              aes(y = estimate,  x=births_total, color="orig. rural"),  size = 1) +
    stat_summary(data=dmm_orig[dmm_orig$births_total <19, ],
                 aes(x=births_total, y=prob_twin,
                     color="orig. rural", fill="orig. rural"),
                 alpha=0.5,
                 fun.data=mean_se) +
    geom_ribbon(data=fig2_orig_plot_data,
                aes(y = estimate,  x=births_total, ymin = lwr, ymax = upr,
                    fill="orig. rural"),
                alpha = 0.3) +
    ggplot2::scale_x_continuous(breaks = 1:18) +
    ggplot2::scale_y_continuous(breaks = seq(0,0.03, by=0.005)) +
    ggplot2::coord_cartesian() +
    labs(subtitle = "Model prediction + data mean with SE",
         y="Per-birth twin. prob.",
         x="Maternal total births")
p2 <- fig2_ext_orig + base_theme(larger=8) + scale_color_manual(values=bc) +
  scale_fill_manual(values=bc) + guides(color="none") + labs(fill = "dataset")
```

```
p2
```

```
## Warning: Removed 1 rows containing missing values (geom_segment).
```

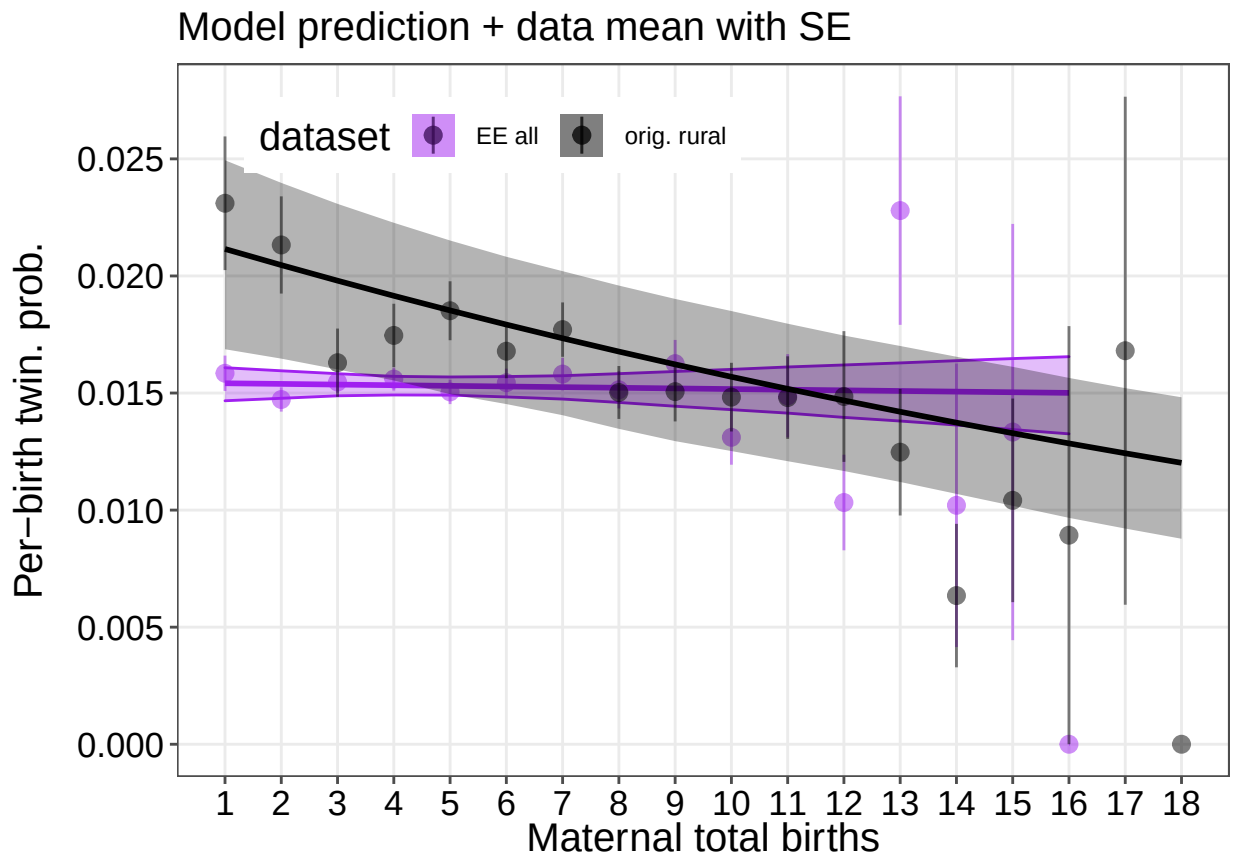


Fig 2: Estonian vs Others No Last Birth

```
fig2_EE_plot_data_nl <- dmm_EE_nl_fit$results
fig2_EE_plot_data_nl$births_total <- fig2_EE_plot_data_nl$births_total + 1
fig2_orig_plot_data_nl <- dmm_orig_nl_fit$results
fig2_orig_plot_data_nl$births_total <- fig2_orig_plot_data_nl$births_total + 1
```

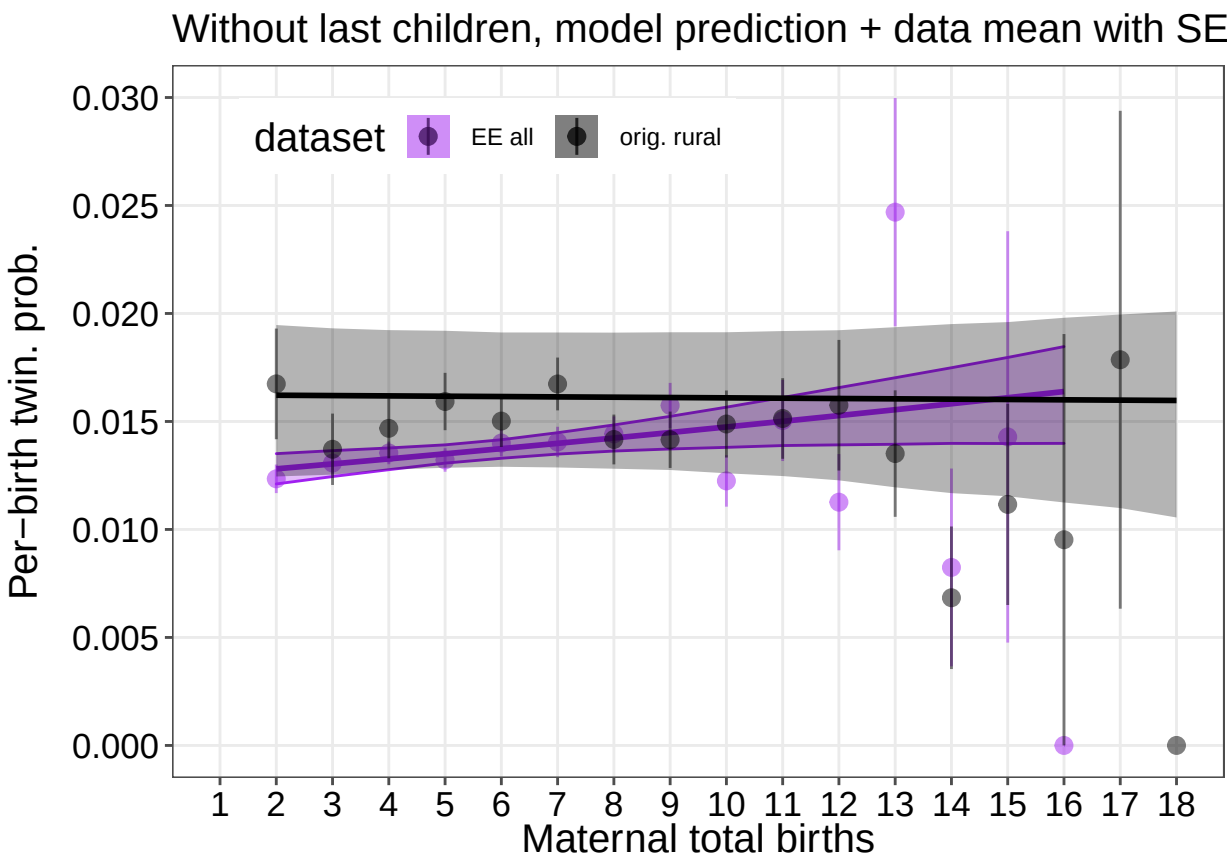
```
dmm_EE_nl_plot <- dmm_EE_nl
dmm_orig_nl_plot <- dmm_orig_nl
dmm_EE_nl_plot$births_total <- dmm_EE_nl_plot$births_total + 1
dmm_orig_nl_plot$births_total <- dmm_orig_nl_plot$births_total + 1
```

```
fig2_ext_orig <- ggplot() +
  geom_line(data=fig2_EE_plot_data_nl,
              aes(y = estimate,  x=births_total, color="EE all"),  size = 1) +
    stat_summary(data=dmm_EE_nl_plot,
                 aes(x=births_total, y=prob_twin, color="EE all", fill = "EE all"),
```

```
                alpha=0.5,
                fun.data=mean_se) +
    geom_ribbon(data=fig2_EE_plot_data_nl,
                aes(y = estimate, x=births_total, ymin = lwr, ymax = upr,
                    color="EE all", fill = "EE all"),
                alpha = 0.3) +
    geom_line(data=fig2_orig_plot_data_nl,
              aes(y = estimate,  x=births_total, color="orig. rural"),  size = 1) +
    stat_summary(data=dmm_orig_nl_plot,
                 aes(x=births_total, y=prob_twin,
                     color="orig. rural", fill="orig. rural"),
                 alpha=0.5,
                 fun.data=mean_se) +
    geom_ribbon(data=fig2_orig_plot_data_nl,
                aes(y = estimate,  x=births_total, ymin = lwr, ymax = upr,
                    fill="orig. rural"),
                alpha = 0.3) +
    ggplot2::scale_x_continuous(breaks = 1:18, limits = c(1,NA)) +
    ggplot2::scale_y_continuous(breaks = seq(0,0.03, by=0.005)) +
    ggplot2::coord_cartesian(ylim=c(0,0.03)) +
    labs(subtitle = "Without last children, model prediction + data mean with SE",
         y="Per-birth twin. prob.",
         x="Maternal total births")
p3 <- fig2_ext_orig + base_theme(larger=8) + scale_color_manual(values=bc) +
  scale_fill_manual(values=bc) + guides(color="none") + labs(fill = "dataset")

p3
```

```
## Warning: Removed 1 rows containing missing values (geom_segment).
```


Without last children, model prediction + data mean with SE
dataset
EE all
orig. rural
Per−birth twin. prob.
0.030
0.025
0.020
0.015
0.010
0.005
0.000
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
Maternal total births

```
#END
```