

# Package ‘childpen’

June 2, 2026

**Title** Identification and Estimation of Child Penalties

**Version** 0.2.3

**Description** Tools to simulate child-penalty data and estimate DID, TD, and NTD identification frameworks from Leventer (2025), “Identification of Child Penalties” <doi:10.48550/arXiv.2602.07486>.

**License** MIT + file LICENSE

**URL** <https://github.com/dorleventer/childpen>,  
<https://dorleventer.github.io/childpen/>

**BugReports** <https://github.com/dorleventer/childpen/issues>

**Imports** data.table, stats

**Suggests** knitr, rmarkdown, testthat (>= 3.0.0), purrr, scales, dplyr,  
tidyr, ggplot2

**Config/testthat/edition** 3

**Encoding** UTF-8

**RoxygenNote** 7.3.3

**VignetteBuilder** knitr

**Depends** R (>= 3.5)

**NeedsCompilation** no

**Author** Dor Leventer [aut, cre]

**Maintainer** Dor Leventer <leventerdor@gmail.com>

**Repository** CRAN

**Date/Publication** 2026-06-02 11:20:07 UTC

## Contents

aggregate_estimands . . . . .	2
multiple_treatment_group_analysis . . . . .	4
simulate_data . . . . .	5
single_treatment_group_analysis . . . . .	6

<b>Index</b>	<b>8</b>
--------------	----------

---

aggregate\_estimands    *Aggregate estimands across treatment groups*

---

### Description

Takes the stacked output of `multiple_treatment_group_analysis()` and computes three aggregate estimands across treatment groups for each event time:

### Usage

```
aggregate_estimands(
  results,
  weights = "sample",
  methods = c("DID_Female", "DID_Male", "TD", "NTD_Conv", "NTD_New"),
  include_pre = FALSE
)
```

### Arguments

results	A data.frame as returned by <code>multiple_treatment_group_analysis()</code> , with at minimum the columns <code>d</code> , <code>event_time</code> , <code>estimand</code> , <code>method</code> , <code>est</code> , and <code>se</code> . If <code>results</code> carries influence-function data (attached automatically by <code>multiple_treatment_group_analysis()</code> ), standard errors account for shared control groups across treatment groups.
weights	How to weight treatment groups. One of: <ul style="list-style-type: none"> <li>"sample" (default): use sample-proportion weights as in Leventer (2025). Within-gender weights <math>w_{g,d} = n_{g,d} / \sum n_{g,\bar{d}}</math> are used for DID_Female and DID_Male; cross-gender weights <math>w_d = n_d / \sum n_{\bar{d}}</math> for TD, NTD_Conv, and NTD_New. Standard errors account for estimation of the weights.</li> <li>NULL: uniform weights (equal weight per group).</li> <li>A named numeric vector: custom fixed weights (names = treatment groups as characters). Values are renormalised to sum to 1.</li> </ul>
methods	Character vector of methods to aggregate. Defaults to all five main methods.
include_pre	Logical. If TRUE, also aggregate pre-treatment event times ( <code>event_time &lt; 0</code> ). Default FALSE.

### Details

**avg\_of\_ratios** ( $\theta_{\text{Agg},1}$ ) Weighted average of the group-specific normalised effects  $\theta(g, d, d + e)$  across treatment groups  $d$ . This is the preferred estimand because it averages effects that are already scaled by each group's baseline.

**ratio\_of\_avgs** ( $\theta_{\text{Agg},2}$ ) Ratio of the weighted-average ATE to the weighted-average APO. The implicit weight on each group is  $p_d \cdot \text{APO}_d$ , giving higher-earning groups more influence.

**gender\_ineq** ( $\Delta\rho_{\text{Agg}}$ ) Weighted average of NTD\_New (estimand == "Delta\_rho") across treatment groups – the aggregate gender-inequality estimand.

**Standard errors.** When the results object carries influence-function (IF) data from `multiple_treatment_group_analysis`, aggregate SEs account for dependence across treatment groups caused by shared control individuals.

With `weights = "sample"`, the IF additionally accounts for estimation of the weights, following the formula in Leventer (2025, Appendix G):

$$\psi_{A(e)} = \sum_d \left[ w_d \psi_{B_d} + \frac{B_d - A(e)}{M} \psi_{p_d} \right]$$

where  $M = \sum_d p_d$  and  $\psi_{p_d}$  is the IF of the group proportion.

With fixed weights (NULL or a named vector), the second term drops out and the IF reduces to  $\sum_d w_d \psi_{B_d}$ .

For `ratio_of_avgs`, the delta method is applied to the ratio  $\bar{\mu}_{ATE}/\bar{\mu}_{APO}$  using the aggregate IFs for the numerator and denominator.

If IF data is not available (e.g., when the user supplies a manually constructed results table), SEs are computed under an independence approximation with a warning.

**Handling missing cells.** Not every treatment group produces an estimate for every event time (due to `max_age / min_age` bounds). The function operates on whichever groups are present for each cell and reports how many via `n_groups`. If `weights` is supplied as a named vector, only the entries whose names appear in the observed treatment groups are used; the remaining weights are dropped and the retained weights are renormalised.

## Value

A data frame with one row per `event_time` by estimand by method by `agg_type` combination, containing:

- `event_time` – event time
- `estimand` – "APO", "ATE", "theta", or "Delta\_rho"
- `method` – method name
- `agg_type` – one of "avg\_of\_ratios", "ratio\_of\_avgs", "gender\_ineq"
- `est` – aggregate estimate
- `se` – standard error (see Details)
- `ci_l`, `ci_h` – 95 %
- `n_groups` – number of treatment groups contributing

## Examples

```
set.seed(1)
sim <- simulate_data(n_individuals = 500)
res <- multiple_treatment_group_analysis(sim, treatment_groups = 24:25,
                                       periods_post = 2, verbose = FALSE)
agg <- aggregate_estimands(res)
head(agg)
```

---

multiple\_treatment\_group\_analysis

*Child penalty analysis over multiple treatment groups*


---

## Description

Child penalty analysis over multiple treatment groups

## Usage

```
multiple_treatment_group_analysis(
  data,
  treatment_groups,
  periods_post,
  periods_pre = 4,
  max_age = 999,
  min_age = 0,
  pre = 1,
  Y_name = "Y",
  age_name = "age",
  D_name = "D",
  id_name = "id",
  female_name = "female",
  verbose = TRUE
)
```

## Arguments

data	A data.frame or data.table with the needed columns. Names can be mapped via Y_name, age_name, D_name, id_name, female_name.
treatment_groups	Integer vector of treatment groups (e.g., 24:34).
periods_post	Integer $H \geq 0$ . Post-treatment horizons; evaluates event times $e = 0, 1, \dots, H$ with target age $a = d + e$ and control $dp = a + 1$ .
periods_pre	Integer $K \geq 0$ (default 4). Number of pre-treatment horizons. Evaluates $e = -K, \dots, -pre$ with $a = d + e$ . For each pre period, tests the same control offsets used post, i.e., $dp = d + 1, 2, \dots, H + 1$ . Set NULL to skip pre-trends.
max_age	Integer (default 999). Upper bound; cells with $dp > max\_age$ are skipped.
min_age	Integer (default 0). Lower bound; cells with $a < min\_age$ are skipped.
pre	Integer (default 1). Pre-treatment anchor used in APO (uses $d - pre$ ).
Y_name, age_name, D_name, id_name, female_name	Column name mappings passed to prep_data_table().
verbose	Logical (default TRUE). Print progress messages.

**Value**

A data.frame stacking results from `single_treatment_group_analysis()`.

**Examples**

```
set.seed(1)
sim <- simulate_data(n_individuals = 500)
res <- multiple_treatment_group_analysis(sim, treatment_groups = 24:25, periods_post = 2,
                                       verbose = FALSE)
head(res)
```

---

 simulate\_data

*Simulate panel data for child-penalty estimation*


---

**Description**

Generates a balanced panel with lifecycle earnings, a gender gap, selection on treatment timing, and gendered treatment effects. The DGP is:

**Usage**

```
simulate_data(n_individuals = 10000, treatment_groups = 24:28, seed = 42)
```

**Arguments**

`n_individuals` Integer. Number of individuals (default 10 000).

`treatment_groups` Integer vector. Treatment groups to include (default 24:28).

`seed` Integer or NULL. RNG seed (default 42). The caller's RNG state is saved and restored on exit, so calling this function does not alter the global random stream. Set to NULL to draw from the current RNG state without reseeding.

**Details**

$$\log Y_{it} = \mu_0 + \lambda D_i + \alpha_i + \beta_1(a-20) + \beta_2(a-20)^2 + \gamma \cdot \mathbf{1}[f] + \theta_f \cdot \mathbf{1}[f, a \geq D] + \theta_m \cdot \mathbf{1}[m, a \geq D] + \varepsilon_{it}$$

where  $\alpha_i \sim N(0, \sigma_\alpha^2)$  is a permanent individual effect and  $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$  is a transitory shock. The term  $\lambda D_i$  generates positive selection on treatment timing: individuals who have children later earn more, on average, than those who have children earlier.

**Value**

A data.frame with columns `id`, `female`, `age`, `D`, `Y`.

**Examples**

```
sim <- simulate_data(n_individuals = 2000)
head(sim)
```

---

```
single_treatment_group_analysis
```

*Unconditional estimands for a single treatment group*

---

**Description**

Estimates 15 descriptive estimands for triplet (treatment group, control group and target age). SEs are calculated using influence-function (IF) calculations with clustering within id.

**Usage**

```
single_treatment_group_analysis(
  data,
  d,
  dp,
  a,
  pre = 1,
  Y_name = "Y",
  age_name = "age",
  D_name = "D",
  id_name = "id",
  female_name = "female"
)
```

**Arguments**

data	A data.table with columns: <ul style="list-style-type: none"> <li>• id — cluster identifier (i.e., person)</li> <li>• age — integer age</li> <li>• female — 0/1 indicator (1 = females)</li> <li>• D — treatment group</li> <li>• Y — numeric outcome</li> </ul>
d	Integer. Treatment group (age at first childbirth)
dp	Integer. Control group (closest not-yet-treated group)
a	Integer. Target age.
pre	Integer, default 1. Offset used for the pre-treatment anchor.
Y_name, age_name, D_name, id_name, female_name	Column name mappings passed to prep_data_table().

## Details

Let  $Y(a, g, d^*)$  denote the mean outcome at age  $a$  for gender  $g \in \{0, 1\}$  (1 = female) when assigned to group  $d^*$ . The core components are:

- $\mathbf{APO}(g; d, d', a) = Y(d - \text{pre}, g, d) + Y(a, g, d') - Y(d - \text{pre}, g, d')$
- $\mathbf{ATE}(g; d, d', a) = Y(a, g, d) - \mathbf{APO}(g; d, d', a)$
- $\mathbf{\theta}(g) = \mathbf{ATE}(g) / \mathbf{APO}(g)$

From these, the cross-gender contrasts are formed:

- $\mathbf{TD} = \mathbf{ATE}(F) - \mathbf{ATE}(M)$
- $\mathbf{NTD\_Conv} = \theta(F) - \theta(M)$
- $\mathbf{NTD\_New} = \frac{Y(a, F, d)}{Y(a, M, d)} - \frac{\mathbf{APO}(F)}{\mathbf{APO}(M)}$
- $\mathbf{TD\_Null}$  and  $\mathbf{NTD\_Conv\_Null}$  variants are defined analogously under a null-effect-for-fathers bias-correction.

Internally, influence functions for all pieces are written into temporary columns of a `data.table` via `compute_mean_if()`, and cluster-robust standard errors are computed by summing the IFs at the `id` level via `se_cluster()`.

## Value

A `data.frame` with one row per estimand/method combination:

- `estimand` — one of "APO", "ATE", "theta", "Delta\_rho"
- `method` — one of "DID\_Female", "DID\_Male", "TD", "NTD\_Conv", "NTD\_New", "TD\_Null", "NTD\_Conv\_Null"
- `est` — estimate
- `se` — cluster-robust standard error
- `n_female_treat`, `n_female_control`, `n_male_treat`, `n_male_control` — sample counts

## Note

Requires helper functions `compute_mean_if()` and `se_cluster()`.

## Examples

```
set.seed(1)
sim <- simulate_data(n_individuals = 500)
res <- single_treatment_group_analysis(sim, d = 25, dp = 26, a = 26, pre = 1)
head(res)
```

# Index

`aggregate_estimands`, [2](#)

`multiple_treatment_group_analysis`, [4](#)

`simulate_data`, [5](#)

`single_treatment_group_analysis`, [6](#)