Estimation of two-stage models in individual participant data with (or without) missing data

Robert Thiesmeier

Dept. Global Public Health, Karolinska Institutet

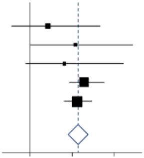
Northern European Stata Conference, Stockholm 2023

Outline

- IPD Meta-analysis (IPDMA)
- `twostage`
- Motivating example: TPC study
- `twostage_post`
- Missing data in IPDMA using `twostage`
- Final remarks

Individual Participant Data Meta Analysis (IPDMA)

- Use of raw data from studies.
- Investigate more granule effects such as treatment covariate interactions and/or joint effects.
- Ensures that similar or same model is fitted in all studies.
- Ensures that relevant participants are included.
- Overcome methodological challenges such as missing data.
- Two-stage approach: 1. Fit the model in each study individually. 2. Combine the estimates from the study.



Options in Stata

Available options for two-stage IPDMA in Stata:

1. `mvmeta` (Ian White), 2009
2. `ipdmetan` (David Fisher), 2013

```
use "ipddata.dta", clear
stset t, fail(failure)
```

mvmeta

```
mvmeta_make stcox trt, strata(sex) by(trialid) ///
names(b V) clear
mvmeta b V, reml
```

idpmetan

```
ipdmetan, study(trialid) re(reml) poolvar(trt) nograph:
///
stcox trt, strata(sex)
```

twostage

```
twostage stcox trt, strata(sex) id(trialid) mv(reml)
```

twostage: facilitating multivariate IPDMA

twostage *command varlist* [if] [in], [cmd options] **[options]**

- twostage works with official and user-written commands.
- Use of **meta mvregress** for multivariate IPDMA.
- Allows for new factor notations **i.** and **#**.
- Allows to be used with **mi estimate**.
- **frames** to allow user to access study specific estimates.
- Access command-specific post-estimation commands.
- Integration of **twostage_postest** for post estimation commands: use with lincom, test, visualisation.

Motivating example: TPC study

Training **P**erformance in **C**ycling (TPC) study

The aim of the study is to assess the joint effect between different training methods and levels of altitude on performance in professional cyclists.

Motivating example: TPC study

- Observational study with around 2000 participants per study and 5 study sites.
- **Training methods**: Aerobic training, VO2 max training, and threshold oriented training.
- **Training altitude**:

$$< 1000, 1000 - 1800m, > 1800m$$

- **Outcome**: Placing in top 5 in professional races (yes/no).
- A single study would have low power to detect any joint effects of training method and altitude. Pooled estimates from 5 study locations increase statistical power.

Motivating example: TPC study

```
tabulate altitude method
           |               method
  altitude |   Aerobic        VO2  Threshold |     Total
-----------+---------------------------------+----------
     <1000 |     1,535      1,137      1,768 |     4,440
 1000-1800 |     1,048        795      1,207 |     3,050
     >1800 |       901        632        977 |     2,510
-----------+---------------------------------+----------
     Total |     3,484      2,564      3,952 |    10,000
```

We are interested in the joint effect of different levels of **altitude** and **training methods**.

Outcome model with indicator variables

$$
\begin{gathered}
logit(P(\text{Winning} = 1 | \text{Training Method, Altitude})) = \\
\beta_0 + \beta_1 \text{ vo2} + \beta_2 \text{ thres} + \beta_3 \text{ med\_alt} + \\
\beta_4 \text{ high\_alt} + \beta_5 \text{ vo2\_med\_alt} + \beta_6 \text{ vo2\_high\_alt} + \\
\beta_7 \text{ thres\_med\_alt} + \beta_8 \text{ thres\_high\_alt}
\end{gathered}
$$

Facilitating twostage

```
use tpc.dta, clear

twostage logit winning vo2 thres med_alt high_alt ///
vo2_med_alt vo2_high_alt thres_med_alt thres_high_alt, ///
id(id) frm(tpc_study_est)

// Alternatively
twostage logit winning i.meth##i.alt, id(id) frm(tpc_study_est)
```

`twostage` gives the following output with the pooled estimates from all variables:

```
Two-stage IPD meta-analysis

Number of studies: 5
Number of participants: 10000
------------------------------------------------------------------------------
              | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
--------------+---------------------------------------------------------------
winning       |
          vo2 |    .053065    .128468     0.41   0.680    -.1987278    .3048577
        thres |   .2261234   .1111222     2.03   0.042     .0083278     .443919
      med_alt |   .1243019   .1290233     0.96   0.335    -.1285792    .3771829
     high_alt |   .0954654   .1377031     0.69   0.488    -.1744277    .3653586
  vo2_med_alt |   .4137719   .1869495     2.21   0.027     .0473576    .7801862
 vo2_high_alt |   .7736476   .1939756     3.99   0.000     .3934625    1.153833
thres_med_alt |   .7589502   .1619021     4.69   0.000     .4416278    1.076273
thres_high_alt |  1.030928   .1703431     6.05   0.000     .6970621    1.364795
        _cons |  -2.171521   .0845473   -25.68   0.000    -2.337231   -2.005812
------------------------------------------------------------------------------
```

Study-specific estimates and SE are saved in a frame.

frame tpc_study_est: summarize

Outline

twostage_post

twostage_post offers post estimation commands for `twostage`.

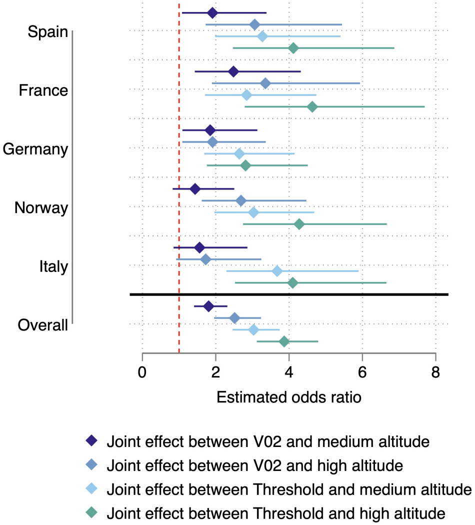
To get linear combinations for the four joint effects of interest we can type:

```
twostage_post ///
    (vo2 + med_alt + vo2_med_alt) ///
    (vo2 + high_alt + vo2_high_alt) ///
    (thres + med_alt + thres_med_alt) ///
    (thres + high_alt + thres_high_alt) , ///
    eform graph
```

twostage_post

```
Twostage post estimation
Linear Combination of vo2 + med_alt + vo2_med_alt

             | Coef   SE     p-val  LB     UB
-------------+------------------------------------
       Spain | 1.911  0.557  0.026  1.079  3.383
      France | 2.483  0.701  0.001  1.428  4.319
     Germany | 1.849  0.498  0.022  1.091  3.134
      Norway | 1.437  0.407  0.200  0.825  2.505
       Italy | 1.559  0.483  0.152  0.849  2.863
-------------+------------------------------------
     Overall | 1.806  0.229  0.000  1.408  2.316
```

twostage_post



With the provided results we can build the final table:

Table 1: Overall joint effects between training method and levels of altitude on cycling performance.

Training Altitude	Training Method Aerobic	VO2	Threshold
<1000	ref.	1.05(0.82,1.36)	1.25(1.01,1.59)
1000-1800	1.13(0.88,1.46)	1.81(1.40,2.32)	3.03(2.46,3.74)
>1800	1.10(0.84,1.44)	2.51(1.96,3.23)	3.87(3.12,4.79)

Conclusion

Threshold training at **medium** and **high altitude** is the most effective training to win cycling races.

Outline

- IPD Meta-analysis (IPDMA)
- twostage
- Motivating example: TPC study
- twostage_post
- Missing data in IPDMA using twostage
- Final remarks

Integration into mi environment

- Systematically missing data in IPDMA can pose practical and methodological challenges for researchers.
- Multiple imputation can be used to learn from other studies and impute values for systematically missing data.
- `mi estimate` currently has no two-stage command.
- `mvmeta` and `ipdmetan` can be used with prefix.

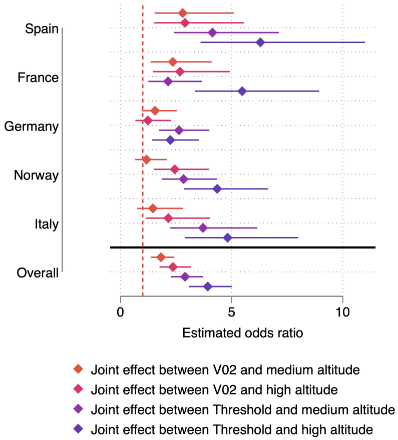
Integration into **mi** `environment`

Let us assume we have some sporadically missing data.

```
use tpcdata.dta, clear

replace altitude = . if runiform()<.4

mi set wide
(...)

mi impute mlogit altitude vo2 thres winning vo2_win thres_win, by(id) add(40)
(...)

mi estimate, dots post: ///
twostage logit winning vo2 thres med_alt high_alt ///
vo2_med_alt vo2_high_alt thres_med_alt thres_high_alt, ///
id(id) frm(tpc_mi_est)
```

```
Multiple-imputation estimates                  Imputations       =         40
Two-stage IPD meta-analysis                    Number of obs     =     10,000
                                               Average RVI       =     0.3824
                                               Largest FMI       =     0.3606
DF adjustment:   Large sample                  DF:     min       =     306.92
                                                       avg       =     623.63
                                                       max       =   1,425.27
Model F test:       Equal FMI                  F(   8, 3417.4)   =      31.18
                                               Prob > F          =     0.0000
--------------------------------------------------------------------------------
               | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
---------------+----------------------------------------------------------------
           vo2 |   .0270338   .1502546     0.18   0.857    -.2680766    .3221441
         thres |   .1634354   .1261954     1.30   0.196    -.0842159    .4110867
       med_alt |   .0057517   .1560645     0.04   0.971    -.3009186     .312422
      high_alt |   .1081341   .1564237     0.69   0.490    -.1990948    .4153631
   vo2_med_alt |   .5649013   .2291951     2.46   0.014     .1142125     1.01559
  vo2_high_alt |   .7226173   .2404113     3.01   0.003     .2495543     1.19568
 thres_med_alt |   .8962267   .1974459     4.54   0.000     .5081353    1.284318
thres_high_alt |   1.095521   .1980394     5.53   0.000     .7064293    1.484612
         _cons |  -2.139291    .093498   -22.88   0.000    -2.322699   -1.955882
--------------------------------------------------------------------------------
```

We can use `twostage_post` again to initiate overall and study specific linear combinations and visualisations.



twostage and mi

Table 1: Overall joint effects between training method and levels of altitude on cycling performance.

Training Altitude	Aerobic	VO2	Threshold
<1000	ref.	1.02(0.78,1.38)	1.17(0.92,1.51)
1000-1800	1.01(0.74,1.36)	1.82(1.37,2.42)	2.90(2.27,3.71)
>1800	1.11(0.82,1.51)	2.36(1.75,3.17)	3.92(3.08,5.00)

Conclusion

No change in the overall conclusion. The values after MI are close to the set parameters.

Final remarks

- twostage offers flexible ways to estimate two-stage models for multivariate IPDMA.
- Possibility for users compute a range of post estimation commands with twostage_post.
- Integration into the mi environment.
- Useful alternative and extension to established commands mvmeta and ipdmetan.

Further readings

- Fisher DJ. *Two-stage individual participant data meta-analysis and generalized forest plots*. Stata Journal 2015;15(2):369-96
- Jolani S, Debray TPA, Koffijberg H, va Buuren S, Moons KGM. *Multiple imputation of systematically missing predictors in an individual participant data meta-analysis: a generalised approach using MICE.* Stat Med 2015; 34(11):1841-1863.
- Morris TP, Fisher DJ, Kenward MG, Carpenter JR. *Meta-analysis of Gaussian individual patient data: Two-stage or not two-stage?* Stat Med, 2018; 37(9):1419-1438.
- Resche-Rigon M, White IR, Bartlett JW, Peters SA, Thompson SG. *Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data.* Stat Med 2013; 32(28):4890-4905.
- Riley RD, Tierney J, Stewart LA (Eds). *Individual Participant Data Meta-Analysis: A Handbook for Healthcare Research*. Chichester: Wiley, 2021.
- Riley RD, et al. *Multivariate meta-analysis using individual participant data.* Res Synth Method. 2015; 6 157-74.
- White, I. R. *Multivariate Random-effects Meta-analysis.* Stata Journal 2009;9(1): 40–56.

robert.thiesmeier@ki.se
Joint work with Nicola Orsini.