

Panel Data and Multilevel Analyses for Academic Publishing Success: Supplemental Handout

Institute for Political Methodology, Taiwan, July 16, 2018

Richard Williams, University of Notre Dame

Lutz Bornmann, Max Plank Society

Andreas Thor, University of Applied Sciences

Generalized Linear Mixed Models (Equations and explanations are copied or adapted from Schunk & Perales, *Stata Journal* 2017(1), pp. 89-115, “Within- and between-cluster effects in generalized linear mixed models: A discussion of approaches and the xthybrid command”)

Basic Model

$$g\{E(y_{ij}|x_{ij}, c_i, u_i)\} = g(\mu_{ij}) = \beta x_{ij} + \gamma c_i + u_i$$

- $g(\cdot)$ is the so called link function. The dependent variable is some sort of function of $E(y)$
 - For linear models, it is often called the identity link. $E(y)$ (the expected value of y) is estimated by the model.
 - For logistic regression it is the logit link. The dependent variable is actually the log odds of success.
- x_{ij} and y_{ij} can have different values across individuals and across clusters e.g. student grades and family income.
- c_i only differs across clusters, e.g. schools can be public or private, but within a school all students are attending either a private school or a public one.
- The error term u_i may reflect level 2 (cluster or group) variables not included in the model. It is assumed to be uncorrelated with the variables that are in the model. If this assumption is violated, there will be omitted variable bias and coefficients will be biased.
- Schunk and Perales assume that omitted variables only occur at level 2. If there are omitted variables at level 1 (e.g. income at the time of the survey) an ε_{ij} term can be added.

Hybrid Model

$$g(\mu_{ij}) = \beta_W(x_{ij} - \bar{x}_i) + \beta_B \bar{x}_i + \gamma c_i + u_i$$

Random Slopes Model

$$g(\mu_{ij}) = (\beta + u_{i2})x_{ij} + \gamma c_i + u_{i1}$$

Analyses

A fixed effects model for highest ranked paper

```
. xtlogit paprbest  nauthors npages nrefs jifperc careerstage i.female i.usa
i.socialscience, nolog fe
note: 799 groups (799 obs) dropped because of all positive or
      all negative outcomes.
note: 1.female omitted because of no within-group variance.
note: 1.usa omitted because of no within-group variance.
note: 1.socialscience omitted because of no within-group variance.
Conditional fixed-effects logistic regression   Number of obs   =   373,535
Group variable: id                            Number of groups  =   13,330
                                              Obs per group:
                                              min =           2
                                              avg =          28.0
                                              max =           683
LR chi2(5)                                    =   7158.64
Prob > chi2                                    =   0.0000
Log likelihood = -33211.053
```

paprbest	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
nauthors	.1020616	.0030442	33.53	0.000	.0960951	.108028
npages	.0046791	.0012776	3.66	0.000	.002175	.0071832
nrefs	.0049528	.000256	19.35	0.000	.0044511	.0054546
jifperc	.0386638	.0006544	59.09	0.000	.0373812	.0399463
careerstage	-.0576846	.0028195	-20.46	0.000	-.0632107	-.0521586
female	0 (omitted)					
Female	0 (omitted)					
usa	0 (omitted)					
Yes	0 (omitted)					
socialscience	0 (omitted)					
Yes	0 (omitted)					

Highest Ranked paper – Fixed Effects Model with Interactions

Variable	Main effects	Female Intr	USA intr	SocSci intr
# authors	0.1020***	-0.0109	0.0103	0.0107
# pages	0.00471***			
# refs	0.00494***			
JIF percentile	0.0400***	0.0007	0.0001	-0.0096***
Career Stage	-0.0653***	0.0208**	0.0200*	0.0109

Random effects model for highly ranked (top quartile) paper

```
. xtlogit topq nauthors npages nrefs jifperc careerstage i.female i.usa
i.socialscience, nolog re
```

```
Random-effects logistic regression      Number of obs      =    374,334
Group variable: id                    Number of groups   =     14,129

Random effects u_i ~ Gaussian          Obs per group:
                                      min =             1
                                      avg =            26.5
                                      max =            683

Integration method: mvaghermite       Integration pts.   =          12

Log likelihood = -209470.86            Wald chi2(8)      =   36801.57
                                      Prob > chi2       =     0.0000
```

topq	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
nauthors	.0729849	.0013875	52.60	0.000	.0702656 .0757043
npages	.014749	.0009566	15.42	0.000	.012874 .016624
nrefs	.006277	.0001762	35.63	0.000	.0059316 .0066223
jifperc	.0399781	.0002369	168.75	0.000	.0395138 .0404425
careerstage	-.0394571	.0008703	-45.34	0.000	-.0411628 -.0377514
female					
Female	-.2309206	.0204889	-11.27	0.000	-.271078 -.1907632
usa					
Yes	.2643758	.0267614	9.88	0.000	.2119244 .3168272
socialscience					
Yes	.1206236	.0276417	4.36	0.000	.0664469 .1748004
_cons	-3.993144	.0237151	-168.38	0.000	-4.039625 -3.946663
/lnsig2u	-.6285584	.0224981			-.6726538 -.584463
sigma_u	.7303151	.0082153			.7143895 .7465957
rho	.1395052	.0027007			.1342955 .1448832

LR test of rho=0: chibar2(01) = 1.6e+04 Prob >= chibar2 = 0.000

Hybrid Model for highly ranked (top quartile) paper

```
. xthybrid topq female usa socialscience, use(nauthors npages nrefs jifperc
careerstage) ///
```

```
> family(binomial) link(logit) clusterid(id) star
```

```
Hybrid model. Family: binomial. Link: logit.
```

Variable	model
topq	
R__female	-0.1694***
R__socialscience	0.0924***
R__usa	0.1860***
W__nauthors	0.0824***
W__npages	0.0106***
W__nrefs	0.0068***
W__jifperc	0.0388***
W__careerstage	-0.0534***
B__nauthors	0.0207***
B__npages	0.0428***
B__nrefs	0.0071***
B__jifperc	0.0518***
B__careerstage	-0.0063***
_cons	-5.1434***
var(_cons[id])	
_cons	0.4649***
Statistics	
ll	-2.088e+05
chi2	37783.1969
p	0.0000
aic	4.177e+05
bic	4.178e+05

legend: * p<.05; ** p<.01; *** p<.001

Level 1: 374334 units. Level 2: 14129 units.

Random Slopes Model for highly ranked (top quartile) paper (10% sample)

```
. melogit topq nauthors npages nrefs jifperc careerstage i.female i.usa
i.socialscience if sample10 || id: careerstage, nolog
Mixed-effects logistic regression      Number of obs   =   37,391
Group variable: id                    Number of groups =    1,413
```

topq	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
nauthors	.0707621	.0044313	15.97	0.000	.0620768	.0794474
npages	.0222111	.0031477	7.06	0.000	.0160418	.0283804
nrefs	.0053515	.0005671	9.44	0.000	.0042399	.006463
jifperc	.039661	.0007399	53.60	0.000	.0382108	.0411112
careerstage	-.0413714	.003255	-12.71	0.000	-.047751	-.0349918
female						
Female	-.1289583	.0633318	-2.04	0.042	-.2530862	-.0048303
usa						
Yes	.3708106	.0844497	4.39	0.000	.2052922	.5363291
socialscience						
Yes	.2172029	.083411	2.60	0.009	.0537203	.3806855
_cons	-4.02058	.0746733	-53.84	0.000	-4.166937	-3.874223

id						
var(careerstage)	.0008324	.0001953			.0005255	.0013185
var(_cons)	.4294519	.0362121			.3640321	.5066281

LR test vs. logistic model: chi2(2) = 1601.45 Prob > chi2 = 0.0000

. estat sd

topq	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
id						
sd(careerstage)	.0288514	.0033854			.0229238	.0363117
sd(_cons)	.6553258	.0276291			.6033507	.7117781

. predict re1 re2 if e(sample), reffects

(calculating posterior means of random effects)
(using 7 quadrature points)
(336943 missing values generated)

```
. sum re1 re2
```

Variable	Obs	Mean	Std. Dev.	Min	Max
re1	37,391	.0010369	.0145811	-.041775	.0541951
re2	37,391	.1278538	.5251565	-1.351337	2.557217