

Longitudinal Logistic Regression: Breastfeeding of Nepalese Children PART II

- GEE models (marginal, population average) – covered last lab
- Random Intercept models (subject specific)
- Transition models

Scientific Question

Determine whether the breastfeeding of Nepalese children varies with child age and/or sex of child.

Data: Nepal Data (nepal.dta) as modified in Lab 10

Outcome: $Y_{ij}=I(\text{breastfeeding}_{ij})$ for individual i at visit number j

We use visit number as our time.

We prepare our dataset like we did in Lab 10.

```
. use "nepal.dta", clear

*****
** Dataset prep work from lab 10 ****
*****

** drop extra variables **
. drop age2 age3 age4 t2

** drop other variables we're not using in this analysis
. drop wt ht arm day month year died alive mage lit

** gen visit number variable to use as our time variable **
. sort id age
. by id: gen visit=_n
. tab visit

      visit |      Freq.      Percent      Cum.
-----+-----
          1 |          200          20.00         20.00
          2 |          200          20.00         40.00
          3 |          200          20.00         60.00
          4 |          200          20.00         80.00
          5 |          200          20.00        100.00
-----+-----
      Total |         1,000         100.00

. xtset id visit
      panel variable:  id (strongly balanced)
      time variable:  visit, 1 to 5
      delta: 1 unit
```

Drop observations with missing values on our outcome

```
. drop if bf==.
(53 observations deleted)

. xtides
```

```

    id: 1, 2, ..., 200          n =      199
   visit: 1, 2, ..., 5         T =        5
      Delta(visit) = 1 unit
      Span(visit)  = 5 periods
      (id*visit uniquely identifies each observation)

Distribution of T_i:   min      5%      25%      50%      75%      95%      max
                    1         4         5         5         5         5         5

      Freq.  Percent   Cum. | Pattern
-----+-----
      170    85.43    85.43 | 11111
       15     7.54    92.96 | 1111.
        5     2.51    95.48 | 1....
        3     1.51    96.98 | 1.111
        3     1.51    98.49 | 111.1
        1     0.50    98.99 | 1.1..
        1     0.50    99.50 | 1.11.
        1     0.50   100.00 | 111..
-----+-----
      199    100.00           | xxxxxx

```

We create a binary indicator of breastfeeding for our outcome variable.

```

*** combine 2 groups to gen 0-1 indicator of breast feeding ***
. gen bfbin=1 if bf==1|bf==2
(564 missing values generated)

. replace bfbin=0 if bfbin==.
(564 real changes made)

```

For our covariates of interest in the analysis, we create a centered age variable, an indicator for male gender and an interaction between male gender and centered age.

```

** create a centered age variable
. gen agec = age-37.8194

* generate a binary indicator for male gender *
. gen male=(sex==1)
. drop sex

* define centered age and male gender interaction term *
. gen agemale=agec*male

```

Subject-specific models accounting for correlation (Random Intercept)

We run a random intercept model for breastfeeding status (y_{ij}) where we account for correlation of the repeated observations on children by including a random intercept for each child (i) and control for child's age (centered).

$$\begin{aligned} \text{logit}P(y_{ij} = 1 | U_i) &= \beta_0 + \beta_1 \text{agec}_{ij} + U_i \\ &= (\beta_0 + U_i) + \beta_1 \text{agec}_{ij} \end{aligned}$$

$$U_i \sim N(0, \sigma_u^2)$$

(At least) two ways to fit a logistic random intercept model in Stata

1. xtlogit
2. gllamm (will be used a lot in the multilevel modeling class)

xtlogit

We'll fit the xtlogit model first since xtlogit is a more simple command

```

Random-effects logistic regression           Number of obs   =       947
Group variable (i): id                     Number of groups =       199

Random effects u_i ~ Gaussian              Obs per group:  min =         1
                                                avg =         4.8
                                                max =         5

Log likelihood = -192.04379                 Wald chi2(1)    =       151.30
                                                Prob > chi2     =        0.0000

```

bfbin	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
agec	-.3193337	.0259612	-12.30	0.000	-.3702168	-.2684506
_cons	-2.116602	.309378	-6.84	0.000	-2.722972	-1.510233

/lnsig2u	1.729395	.1716709			1.392927	2.065864

sigma_u	2.374288	.2037981			2.006643	2.809291
rho	.6314745	.0399503			.5503486	.7057885

Likelihood-ratio test of rho=0: chibar2(01) = 126.34 Prob >= chibar2 = 0.000						

Report results on the OR scale:

bfbin	OR	Std. Err.	z	P> z	[95% Conf. Interval]	
agec	.726633	.0188643	-12.30	0.000	.6905846	.7645632

/lnsig2u	1.729395	.1716709			1.392927	2.065864

sigma_u	2.374288	.2037981			2.006643	2.809291
rho	.6314745	.0399503			.5503486	.7057885

Coefficient estimates have **subject-specific interpretation!**

Interpretation of OR for agec:

For a given child, the odds of breastfeeding decreases by 28% for a one month increase in that child's age.

If this result was from a GEE (population average model), the interpretation would be:
On average, for the Nepalese children in our study, the odds of breastfeeding decreases by 40% for each one month increase in age.

Intra-class correlation in our model

Two equivalent interpretations:

1. The proportion of the total variance in breastfeeding status that is due to differences between children (i.e., the variance in the random intercept for child) after controlling for age.
2. The correlation of the repeated measurements of breastfeeding status on the same child after controlling for age.

$$rho \approx \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2}$$

In our model, rho = 0.63

The variance of the random intercept is $(\sigma_u)^2 = 5.61$

Digression on using xtlogit – check for adequate fitting of the model

xtlogit uses quadrature methods to approximate the likelihood function since there is no closed form solution. This means xtlogit and other methods that use quadrature can run slowly and we need to check that the specified quadrature has adequately approximated the likelihood.

```
. quadchk, nooutput
```

```
Refitting model intpoints() = 8
Refitting model intpoints() = 16
```

Quadrature check

	Fitted quadrature 12 points	Comparison quadrature 8 points	Comparison quadrature 16 points	
Log likelihood	-180.52843	-180.97225	-180.58976	
		-.44381901	-.06132697	Difference
		.00245844	.00033971	Relative difference
bfbin:	-.51073187	-.55123278	-.50715679	
agec		-.04050091	.00357508	Difference
		.07929975	-.00699991	Relative difference
bfbin:	-3.5069938	-3.7425012	-3.4656029	
_cons		-.23550739	.04139091	Difference
		.06715364	-.01180239	Relative difference
lnsig2u:	3.1247561	3.2928336	3.1086299	
_cons		.16807743	-.01612628	Difference
		.05378898	-.00516081	Relative difference

Relative difference > 10⁻² (1%) (arbitrary STATA rule of thumb) so refit using > # of integration points (default was 12).

```
. xtlogit bfbin agec, i(id) or intpoints(18) nolog
```

Random-effects logistic regression	Number of obs	=	947
Group variable (i): id	Number of groups	=	199
Random effects u_i ~ Gaussian	Obs per group: min	=	1
	avg	=	4.8
	max	=	5
	Wald chi2(1)	=	129.57
Log likelihood = -187.77104	Prob > chi2	=	0.0000

bfbin	OR	Std. Err.	z	P> z	[95% Conf. Interval]
agec	.7013965	.0218547	-11.38	0.000	.6598439 .7455659
/lnsig2u	2.05032	.1814144			1.694754 2.405886
sigma_u	2.787541	.2528501			2.333518 3.329902
rho	.7025504	.0379108			.623377 .7711892

Likelihood-ratio test of rho=0: chibar2(01) = 134.88 Prob >= chibar2 = 0.000

```
. quadchk, nooutput
```

```
Refitting model intpoints() = 14
Refitting model intpoints() = 22
```

Quadrature check

	Fitted quadrature 18 points	Comparison quadrature 14 points	Comparison quadrature 22 points	
Log likelihood	-187.77104	-188.93518	-185.07007	
		-1.1641388	2.7009735	Difference
		.00619978	-.0143844	Relative difference
bfbin:	-.35468188	-.35307272	-.38956901	
agec		.00160917	-.03488712	Difference
		-.00453693	.09836173	Relative difference
bfbin:	-2.372401	-2.347957	-2.6170863	
_cons		.02444394	-.24468535	Difference
		-.01030346	.10313828	Relative difference
lnsig2u:	2.05032	2.0176587	2.3306293	
_cons		-.03266129	.28030933	Difference
		-.01592985	.13671491	Relative difference

We can increase intpoints again for even more assurance that the likelihood is appropriately approximated.

```
. xtlogit bfbin agec, i(id) or intpoints(100) nolog
```

Random-effects logistic regression	Number of obs	=	947
Group variable (i): id	Number of groups	=	199
Random effects u_i ~ Gaussian	Obs per group: min	=	1
	avg	=	4.8

```

max = 5
Wald chi2(1) = 57.02
Prob > chi2 = 0.0000
Log likelihood = -180.73476

```

bfbin	OR	Std. Err.	z	P> z	[95% Conf. Interval]	
agec	.6082483	.0400476	-7.55	0.000	.53461	.6920298
/lnsig2u	3.04669	.3048884			2.44912	3.64426
sigma_u	4.587545	.6993446			3.402668	6.18502
rho	.8648116	.0356453			.7787286	.9208107

```

Likelihood-ratio test of rho=0: chibar2(01) = 148.96 Prob >= chibar2 = 0.000
Likelihood-ratio test of rho=0: chibar2(01) = 149.21 Prob >= chibar2 = 0.000

```

```
. quadchk, nooutput
```

```

Refitting model intpoints() = 67
Refitting model intpoints() = 133

```

Quadrature check

	Fitted quadrature 100 points	Comparison quadrature 67 points	Comparison quadrature 133 points	
Log likelihood	-180.60658	-180.60658	-180.60658	
		2.263e-06	2.245e-07	Difference
		-1.253e-08	-1.243e-09	Relative difference
bfbin:	-.51104731	-.51104799	-.51104788	
agec		-6.763e-07	-5.726e-07	Difference
		1.323e-06	1.120e-06	Relative difference
bfbin:	-3.4892956	-3.489302	-3.4892996	
_cons		-6.446e-06	-4.033e-06	Difference
		1.847e-06	1.156e-06	Relative difference
lnsig2u:	3.1257532	3.125756	3.1257555	
_cons		2.863e-06	2.330e-06	Difference
		9.159e-07	7.454e-07	Relative difference

Note the difference! Now the OR for agec is 0.6 and the estimated heterogeneity is 20.9.

GLLAMM

generalized linear latent and mixed models

(gllamm can do a lot more than just logistic models with a random intercept!)

Install gllamm (you have to be connected to the internet)

```
ssc describe gllamm
```

or update gllamm (replace previous version)

```
ssc install gllamm, replace
```

Fit the same random intercept model where we control for age.

$$\begin{aligned} \text{logit}P(y_{ij} = 1 | U_i) &= \beta_0 + \beta_1 \text{agec}_{ij} + U_i \\ &= (\beta_0 + U_i) + \beta_1 \text{agec}_{ij} \\ U_i &\sim N(0, \sigma_u^2) \end{aligned}$$

```
. gllamm bfbinc agec, i(id) l(logit) f(binom)
```

```
number of level 1 units = 947
number of level 2 units = 199
```

bfbinc	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
agec	-.4812358	.0676832	-7.11	0.000	-.6138925	-.3485791
_cons	-3.183983	.5831388	-5.46	0.000	-4.326914	-2.041052

Variances and covariances of random effects

```
***level 2 (id) var(1): 15.385845 (4.301343)
```

```
** get OR instead of log(OR) **
```

```
. gllamm, eform
```

bfbinc	exp(b)	Std. Err.	z	P> z	[95% Conf. Interval]	
agec	.6180192	.0418295	-7.11	0.000	.54124	.7056901

Variances and covariances of random effects

```
***level 2 (id) var(1): 15.385845 (4.301343)
```

gllamm reports the variance of the random intercept, $\sigma_u^2=15.38$

We'll cover gllamm in much more detail 4th term.

Now, we are going to compare the GEE models (lab 10) to the random intercept model.

First, we obtain the predicted log odds from the random intercept model (gllamm).

```
* get the fitted log odds
. predict fittedLO, xb
(xb will be stored in fittedLO)
```

Next, we assign values to the random intercept for each child using Empirical Bayes estimates (we'll cover this more in 4th term).

```
. gllapred ebRI, u
(means and standard deviations will be stored in ebRIml ebRIs1)
```

Produce a variable containing the fitted probability of breastfeeding for an 'average' child (a child with a value of the random intercept = 0).

```
. gen fitted_avgp = exp(fittedLO)/( 1 + exp(fittedLO) )
```

Produce a variable containing the fitted probability of breastfeeding for each child.

```
. gen fitted_indp = exp(fittedLO + ebRIm1)/( 1 + exp(fittedLO + ebRIm1) )
```

Now we refit GEE models with three correlation structures (from lab 10) and store the fitted values.

```
. quietly xtgee bfbinc agec, f(bin) link(logit) corr(ar1) robust
. predict fitted_ggear1, mu
. label var fitted_ggear1 "gee_ar1"

. ****Fit GEE with uniform corr model ****
. quietly xtgee bfbinc agec, f(bin) link(logit) corr(exc) robust
. predict fitted_ggeunif, mu
. label var fitted_ggeunif "gee_unif"

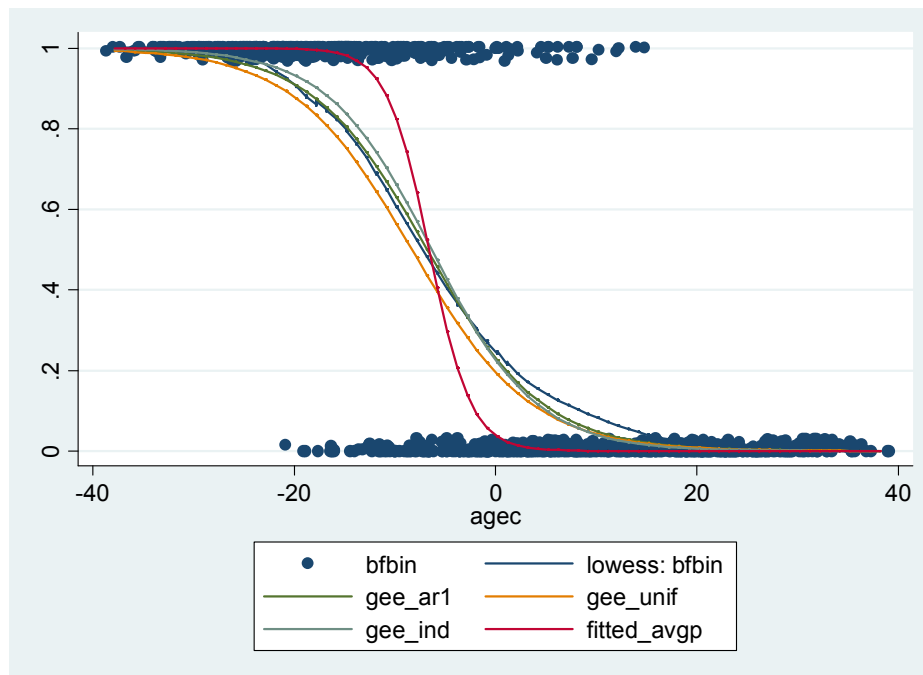
. ****Fit GEE with independence corr model ****
. quietly xtgee bfbinc agec, f(bin) link(logit) corr(ind) robust
. predict fitted_ggeind, mu
. label var fitted_ggeind "gee_ind"
```

Smooth the observed breastfeeding data versus agec

```
. ksm bfbinc agec, lowess bw(.4) ylab(0(.2)1) lwidth(10) gen(bfbincsm)
```

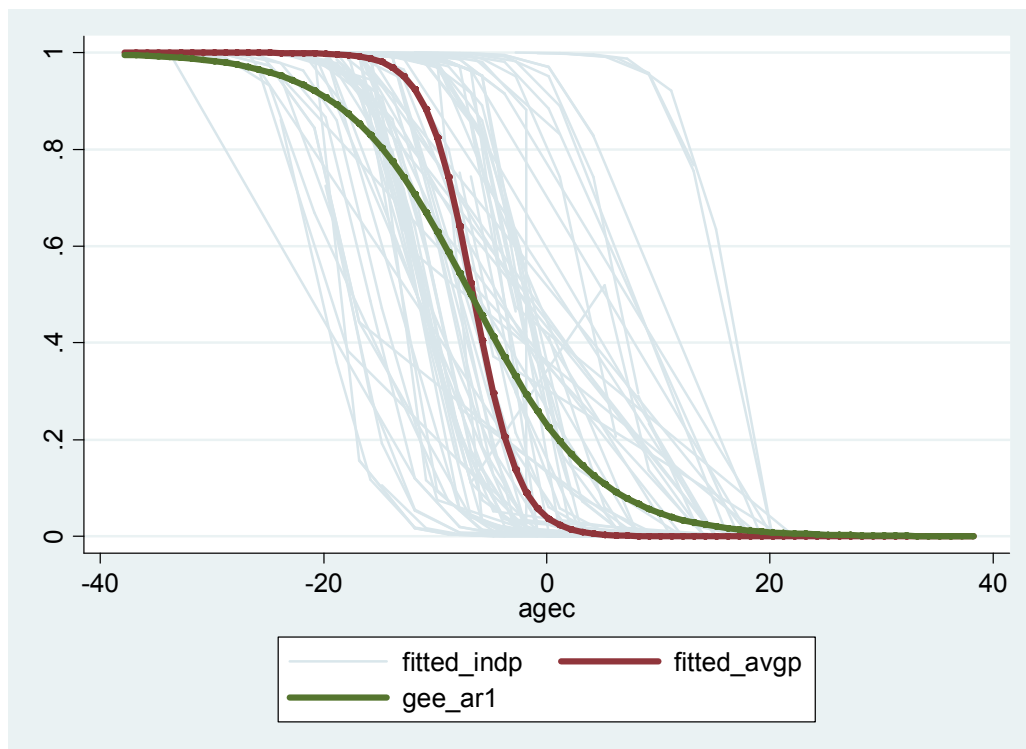
Plot the GEE and RI model fits

```
. twoway (scatter bfbinc agec, jitter(4)) (line bfbincsm fitted_ggear1
fitted_ggeunif fitted_ggeind fitted_avgp agec, sort pstyle (p1)),
ylab(0(.2)1)
```



The fit from random intercept model is quite different from those from marginal model. This is to be expected because they are estimating different things and have different interpretations. The OR from the random intercept model is larger in absolute value (the logistic curve is steeper), as is guaranteed by theory. (Check the book!)

```
twoway (line fitted_indp agec, c(L) pstyle(p15)) (line fitted_avgp agec
, sort clwidth(thick)) (line fitted_gear1 agec , sort clwidth(thick))
```



Transition model

Basic idea: include the outcome variable at a previous time point as a fixed effect covariate. We condition the response at time j on the response at time $j-1$ or $j-2$, etc...

Generate a variable that represents breastfeeding status during the previous month. We often call this lag 1 breastfeeding status.

```
. sort id visit
. by id: gen bfbn_lag1 = bfbn[_n-1]
(199 missing values generated)
```

A very simple transition model

We'll assume *conditional independence* once we condition on the outcome at the previous month (first order Markov chain)

$$\text{logit}P(y_{ij} = 1 | U_i) = \beta_0 + \beta_1 \text{agec}_{ij} + \beta_2 y_{i(j-1)}$$

```
. logit bfbin agec bfbin_lag1
```

```
Logistic regression                Number of obs   =          748
                                LR chi2(2)       =        725.19
                                Prob > chi2        =          0.0000
Log likelihood = -133.01167        Pseudo R2      =          0.7316
```

bfbin	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
agec	-.0976751	.01661	-5.88	0.000	-.13023	-.0651201
bfbin_lag1	4.623251	.5610318	8.24	0.000	3.523649	5.722853
_cons	-3.899484	.507124	-7.69	0.000	-4.893429	-2.905539

```
. logistic bfbin agec bfbin_lag1
```

```
Logistic regression                Number of obs   =          748
                                LR chi2(2)       =        725.19
                                Prob > chi2        =          0.0000
Log likelihood = -133.01167        Pseudo R2      =          0.7316
```

bfbin	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
agec	.9069436	.0150643	-5.88	0.000	.8778935	.936955
bfbin_lag1	101.8245	57.12679	8.24	0.000	33.90793	305.776

Interpretation of OR on agec (population average)

Comparing children of the same breastfeeding status at the previous month, a one month increase in age is associated with a 10% decrease in the odds of breastfeeding.

Interpretation of OR on bfbin_lag1 (population average)

Comparing children of the same age, those children who were breastfeeding at the previous month have an odds ratio of breastfeeding that is 101 times greater than the odds ratio of breastfeeding for those children who were not breastfeeding at the previous month.

Huge!! Note that the 95% CI for the lag 1 bfbin variable does not include 1, so we should keep this variable in our model.

Less simple transition model

We'll allow for a correlation structure on the responses even after we condition on the outcome at the previous month

```
. xtgee bfbin agec bfbin_lag1, nolog f(bin) l(logit) corr(unst) robust

GEE population-averaged model
Group and time vars:      id visit
Link:                     logit
Family:                   binomial
Correlation:              unstructured
Scale parameter:         1
Number of obs             =      748
Number of groups         =      194
Obs per group: min      =         1
                       avg      =      3.9
                       max      =         4
Wald chi2(2)             =     126.98
Prob > chi2              =      0.0000
```

(Std. Err. adjusted for clustering on id)

bfbin	Coef.	Semi-robust Std. Err.	z	P> z	[95% Conf. Interval]	
agec	-.0919938	.01988	-4.63	0.000	-.1309579	-.0530297
bfbin_lag1	5.050811	.7124671	7.09	0.000	3.654401	6.447221
_cons	-4.18535	.6244153	-6.70	0.000	-5.409182	-2.961519

```
. xtcorr
```

Estimated within-id correlation matrix R:

	c1	c2	c3	c4
r1	1.0000			
r2	-0.0261	1.0000		
r3	-0.0952	-0.0990	1.0000	
r4	0.0847	0.0671	-0.1096	1.0000

The working correlation matrix estimates correlations that are all relatively close to zero, We'll keep the unstructured correlation with robust as our final model.

```
. xtgee bfbin agec male agemale bfbin_lag1, nolog f(bin) l(logit)
corr(unst) robust eform
```

```
GEE population-averaged model
Group and time vars:      id visit
Link:                     logit
Family:                   binomial
Correlation:              unstructured
Scale parameter:         1
Number of obs             =      748
Number of groups         =      194
Obs per group: min      =         1
                       avg      =      3.9
                       max      =         4
Wald chi2(4)             =     123.63
Prob > chi2              =      0.0000
```

(Std. Err. adjusted for clustering on id)

bfbin	Odds Ratio	Semi-robust Std. Err.	z	P> z	[95% Conf. Interval]	
agec	.9072332	.0262031	-3.37	0.001	.8573027	.9600718
male	.8001362	.4475089	-0.40	0.690	.2673589	2.394601
agemale	1.01493	.0366755	0.41	0.682	.9455337	1.089419
bfbin_lag1	181.1121	131.0482	7.19	0.000	43.857	747.9211

```
. xtcorr
```

Estimated within-id correlation matrix R:

	c1	c2	c3	c4
r1	1.0000			

```

r2  -0.0312    1.0000
r3  -0.1078   -0.1107    1.0000
r4   0.0936    0.0636   -0.1266    1.0000

```

Test for a gender effect

```
. test male agemale
```

```
( 1) male = 0
```

```
( 2) agemale = 0
```

```

          chi2( 2) =    1.58
Prob > chi2 =    0.4538

```

We still do not see an effect of gender on the breastfeeding status of the Nepalese children.

Final note on the transitions observed in breastfeeding status:

```
. xttrans bfbin
```

bfbin	bfbin		Total
	0	1	
0	99.06	0.94	100.00
1	13.93	86.07	100.00
Total	62.30	37.70	100.00

Maybe our model still doesn't adequately represent the data. There are a class of models called 'mover-stayer' models developed in the 1950's that attempt to represent data of this sort (we don't cover these models in this class). You can find tons of information on these models with a Google search.