

Longitudinal Logistic Regression: Breastfeeding of Nepalese Children

Scientific Question

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

Data: Nepal Data (nepal.dta)

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

We will use the visit number as our time variable (similar to using ‘grouped’ time in the midterm).

First, we change directories and load the data

```
. cd "C:\Documents and Settings\Sandrah Eckel\Desktop\LDA lab10"
C:\Documents and Settings\Sandrah Eckel\Desktop\LDA lab10

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

Next we prepare our data for analysis.

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

** 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
```

We have information on 200 children for 5 visits.

```
. xtdes

      id:  1, 2, ..., 200      n =      200
      visit: 1, 2, ..., 5      T =      5
```

```
Delta(visit) = 1; (5-1)+1 = 5
(id*visit uniquely identifies each observation)
```

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

```

      Freq.  Percent   Cum. | Pattern
-----+-----
      200    100.00  100.00 | 11111
-----+-----
      200    100.00      | XXXXX
```

Our outcome of interest is the breastfeeding status of the child. From the readme file on the nepal.dta dataset, we have a description of our breastfeeding variable bf:

```
bf:
  Indicates current level of breastfeeding:
    0 = none;
    1 = <10 times/day;
    2 = 10 or more times/day.
```

```
. codebook bf
```

```
-----
bf (unlabeled)
-----
```

```

          type:  numeric (byte)
          range:  [0,2]
unique values:  3
          units:  1
          missing.: 53/1000
```

```

tabulation:  Freq.  Value
              564    0
              151    1
              232    2
               53    .
```

We have 53 missing values for breastfeeding. Any of the commands that we will be using in Stata will automatically remove each of these missing observations from the dataset (but retain other non-missing observations for each child with some non-missing bf information). Let's make this explicit in our exploratory data analysis by dropping the observations with missing values for bf.

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

```
. xtides
```

```

      id:  1, 2, ..., 200          n =      199
      visit: 1, 2, ..., 5        T =       5
      Delta(visit) = 1; (5-1)+1 = 5
      (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		XXXXX

Our data are no longer ‘balanced’ or ‘equally spaced’.

We create a binary indicator of breastfeeding status (ever vs. never):

```
. gen bfbin=1 if bf==1|bf==2
(564 missing values generated)

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

. tab bf bfbin
```

bf	bfbin		Total
	0	1	
0	564	0	564
1	0	151	151
2	0	232	232
Total	564	383	947

Let’s explore the distribution of our outcome variable a little more:

```
. xttab bfbin
```

bfbin	Overall		Between		Within
	Freq.	Percent	Freq.	Percent	Percent
0	564	59.56	142	71.36	82.46
1	383	40.44	101	50.75	80.63
Total	947	100.00	243	122.11	81.70

(n = 199)

Interpretation:

Overall, at 59.56% of our child-year observations, we see children that are not breastfeeding (bfbin=0). Taking each child individually, 71.36% of the children are at some point not breastfeeding (bfbin=0), 50.75% of the children are at some point breastfeeding (bfbin=1). Thus, some children are breastfeeding at some visits and not at other visits. Taking children one at a time, if a child is ever not breastfeeding, 82.46% of that child’s observations are not breastfeeding. If a child is ever breastfeeding, 80.63% of

that child's observations are breastfeeding. If breastfeeding status never varied, the within percentages would all be 100%.

```
. xttrans bfbn
```

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

Interpretation:

The top left cell tells us that if a child was not breastfeeding at the previous visit, the probability that the child will not be breastfeeding at the current visit is 0.9906. From the top right cell we see that the probability that a child who was not breastfeeding at the previous visit will be breastfeeding at the current visit is 0.0094. The probability in the bottom left corner, 0.1393, is the probability that a child who was breastfeeding at the previous visit is not breastfeeding at the current visit. Finally, the bottom right corner gives the probability that a child who was breastfeeding at the previous visit is still breastfeeding at the current visit.

Next explore our covariates of interest:

```
. codebook age
```

```
-----
age (unlabeled)
-----
      type:  numeric (byte)
      range:  [0,76]
unique values: 77
      units:  1
      missing.: 0/947
      mean:   37.8194
      std. dev: 18.6181
      percentiles: 10%    25%    50%    75%    90%
                   13     22     37     53     63
-----
```

Create a centered age variable

```
. gen agec = age-37.8194
```

```
. codebook sex
```

```
-----
sex (unlabeled)
-----
      type:  numeric (byte)
      range:  [1,2]
unique values: 2
      units:  1
      missing.: 0/947
-----
```

```

tabulation:  Freq.  Value
              500    1
              447    2

```

```

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

. tab sex male

```

sex	male		Total
	0	1	
1	0	500	500
2	447	0	447
Total	447	500	947

```

. drop sex

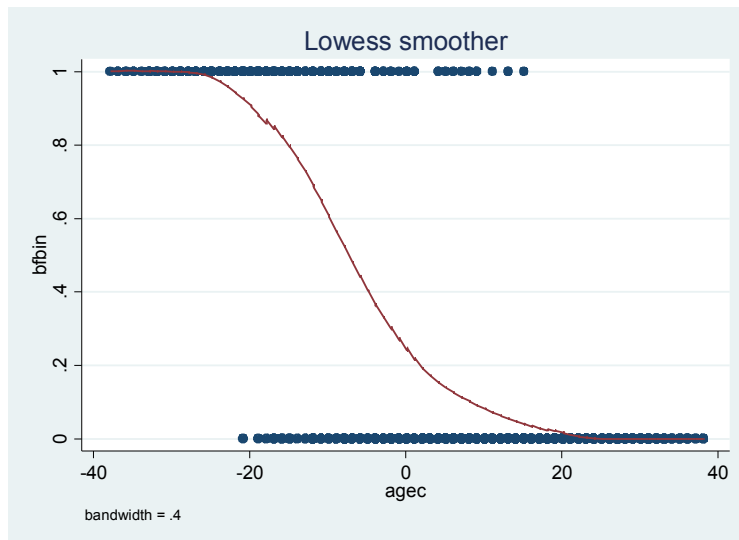
```

Explore the marginal mean model with respect to age

```

. ksm bfbinsm agec, lowess bw(.4) ylab(0(.2)1) gen(bfbinsm)

```

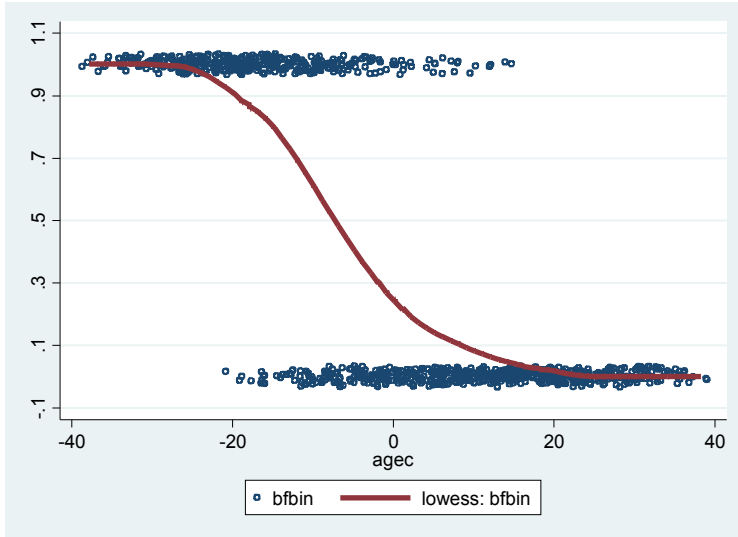


Plot the smooth again, this time with jittered observed outcome and wider smooth line

```

. twoway (scatter bfbinsm agec, jitter(4) msymbol(oh)) (line
bfbinsm agec, sort lwidth(1)), ylab(-.1(.2)1.1)

```



It appears that the logistic function will be appropriate for modeling the effects of child's age on prevalence of bf.

Review of logistic regression in STATA without taking into account correlation of repeated observations on the same children

First, generate an interaction term (agem) between age and male gender

```
. gen agemale=agec*male

. logit bfbin male agec agemale
```

```
Logistic regression                Number of obs   =          947
                                LR chi2(3)       =          770.85
                                Prob > chi2        =           0.0000
Log likelihood = -253.58013        Pseudo R2      =           0.6032
```

bfbin	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
male	.4832121	.2916797	1.66	0.098	-.0884697 1.054894
agec	-.2029085	.0199282	-10.18	0.000	-.2419671 -.1638499
agemale	.0127863	.0263864	0.48	0.628	-.03893 .0645026
_cons	-1.515399	.2331096	-6.50	0.000	-1.972285 -1.058512

Logit reports coefficient estimates on the log-odds scale, logistic reports the coefficient estimates on the odds scale.

```
. logistic bfbin male agec agemale
```

```
Logistic regression                Number of obs   =          947
                                LR chi2(3)       =          770.85
                                Prob > chi2        =           0.0000
Log likelihood = -253.58013        Pseudo R2      =           0.6032
```

bfbin	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
male	1.621274	.4728927	1.66	0.098	.9153309	2.87167
agec	.8163529	.0162685	-10.18	0.000	.785082	.8488694
agemale	1.012868	.0267259	0.48	0.628	.961818	1.066628

Get the predicted probability of breastfeeding from the regression

```
. predict prob
(option p assumed; Pr(bfbin))
```

Create a 2x2 table based on a cutoff (here we choose $c=0.5$) for the predicted probability. We can use this table to calculate the sensitivity and specificity of our predictive model.

```
. gen c = 0.5
. gen bfhat = 1 if prob > c
(553 missing values generated)
. replace bfhat = 0 if bfhat == .
(553 real changes made)
. tab bfbin bfhat
```

bfbin	bfhat		Total
	0	1	
0	504	60	564
1	49	334	383
Total	553	394	947

Test for the statistical significance of the interaction term between age and the male gender

```
. test agemale

( 1) agemale = 0

             chi2( 1) =    0.23
             Prob > chi2 =    0.6280
```

Test whether there is a gender effect

```
. test male agemale

( 1) male = 0
( 2) agemale = 0

             chi2( 2) =    3.19
             Prob > chi2 =    0.2028
```

Don't use these tests of statistical significance! We haven't yet taken the correlation into account so the standard errors on which these tests are based are incorrect!!!

Let's move on to modeling the probability of breastfeeding while taking into account the correlation between repeated observations on the same child.

Note on correlation structure of repeated measures of a binary outcome

We won't explicitly explore the correlation structure like we did for continuous outcomes using the autocorrelation function and variogram. You can explore the correlation structure of binary outcomes using the lorelogram (see p. 52 of Diggle, Heagerty, Liang and Zeger and Heagerty and Zeger (1998)) but, as far as we know, there is no implementation of the lorelogram using STATA. You can find R code for creating lorelograms on the software page of our website.

3) Restrict estimation to a subset of the data that produces balanced panels. (This can help, but it is not guaranteed to produce convergence.)

Remember that correct specification of the correlation structure affects only the efficiency of the parameter estimates. The estimates are consistent regardless of correlation structure. While correct coverage by the default standard error estimates requires a correct correlation structure, this requirement can be relaxed by adding the robust option. **When robust is specified, not only are the parameter estimates consistent, their standard error estimates have correct coverage, regardless of whether the "true" correlation structure is specified.**

So, we can't use a GEE with unstructured correlation for this data.

You can look at the latest estimate of the working correlation matrix (even though we encountered a not positive definite matrix). Keep in mind that this is not a 'final' estimate of the correlation structure. Use it only to get a general sense of where the xtgee model fitting procedure was heading when trying to estimate an unstructured correlation matrix...but nothing more!

```
. xtcorr  
  
Estimated within-id correlation matrix R:  
  
      c1      c2      c3      c4      c5  
r1  1.0000  
r2  0.2439  1.0000  
r3  0.2642  0.4383  1.0000  
r4  0.3181  0.4913  0.6575  1.0000  
r5  0.3205  0.5187  0.7040  1.0000  1.0000
```

We'll try AR-1, a more restrictive model for the correlation structure but that allows the correlation to decrease as the separation between visits increases.

GEE with AR-1 correlation structure

```
. xtgee bfbin male agec agemale, f(bin) link(logit) corr(ar1)
```

**note: observations not equally spaced
modal spacing is delta visit = 1
8 groups omitted from estimation**
**note: some groups have fewer than 2 observations
not possible to estimate correlations for those groups
5 groups omitted from estimation**

```
GEE population-averaged model  
Group and time vars:      id visit      Number of obs      =      913  
Link:                      logit        Number of groups   =      186  
Family:                    binomial    Obs per group: min =        3  
Correlation:               AR(1)        avg =              4.9  
Scale parameter:          1            max =              5  
Wald chi2(3)              =      106.63  
Prob > chi2                =        0.0000
```

```
-----+-----  
      bfbin |      Coef.   Std. Err.      z    P>|z|    [95% Conf. Interval]  
-----+-----  
      male |   .2333108   .4099956    0.57  0.569   - .5702659   1.036887  
      agec |  - .1780666   .0248581   -7.16  0.000   - .2267876  - .1293456
```

```

    agemale |   .0012619   .0344276   0.04   0.971   -.0662149   .0687386
      _cons |  -1.329081   .3094716  -4.29   0.000   -1.935634  -0.7225282
-----+-----

```

```
. xtcorr
```

```

           c1      c2      c3      c4      c5
r1  1.0000
r2  0.6037  1.0000
r3  0.3644  0.6037  1.0000
r4  0.2200  0.3644  0.6037  1.0000
r5  0.1328  0.2200  0.3644  0.6037  1.0000

```

The AR-1 model requires equally spaced data that has an adequate number of observations, so we end up dropping data on a total of 13 children.

GEE with uniform correlation structure

```
. xtgee bfbin male agec agemale, f(bin) link(logit) corr(exc)
```

```

GEE population-averaged model
Group variable:          id      Number of obs      =      947
Link:                   logit    Number of groups   =      199
Family:                 binomial  Obs per group: min =       1
Correlation:           exchangeable  avg               =      4.8
                                           max               =       5
                                           Wald chi2(3)      =     148.30
Scale parameter:       1         Prob > chi2        =      0.0000

```

```

-----+-----
    bfbin |      Coef.   Std. Err.    z    P>|z|    [95% Conf. Interval]
-----+-----
    male |  -0.3312421  .3841878   -0.86  0.389   -1.084236   .4217522
    agec |  -0.1519764  .0185056  -8.21  0.000   -0.1882467  -0.115706
  agemale |  -0.031926   .0275889  -1.16  0.247   -0.0859994   .0221473
    _cons | -1.259992   .2698996  -4.67  0.000   -1.788985   -0.7309982
-----+-----

```

```
. xtcorr
```

```
Estimated within-id correlation matrix R:
```

```

           c1      c2      c3      c4      c5
r1  1.0000
r2  0.5496  1.0000
r3  0.5496  0.5496  1.0000
r4  0.5496  0.5496  0.5496  1.0000
r5  0.5496  0.5496  0.5496  0.5496  1.0000

```

Our estimated coefficients are different comparing the uniform and AR-1 model results. **Why is this?** A GEE model should give consistent estimates of the model coefficients regardless of the correlation structure. One key difference between the two models is that we are estimating the two models using different datasets. The uniform model uses data on 199 children while the AR-1 model uses data on only 186 (13 fewer) children.

GEE with independence correlation structure

```
. xtgee bfbin male agec agemale, f(bin) link(logit) corr(ind)
```

```

GEE population-averaged model
Group variable:          id
Link:                   logit
Family:                 binomial
Correlation:            independent
Scale parameter:        1
Pearson chi2(947):      739.70
Dispersion (Pearson):   .7810953
Number of obs          = 947
Number of groups       = 199
Obs per group: min     = 1
                   avg  = 4.8
                   max  = 5
Wald chi2(3)           = 225.32
Prob > chi2            = 0.0000
Deviance               = 507.16
Dispersion              = .5355441

```

```

-----+-----
      bfbbin |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      male |   .4832121   .2916799     1.66   0.098   - .0884699   1.054894
      agec |  -.2029085   .0199282    -10.18  0.000   - .2419672  -.1638499
    agemale |   .0127863   .0263864     0.48   0.628   - .03893    .0645027
      _cons |  -1.515399   .2331097    -6.50   0.000   -1.972285  -1.058512
-----+-----

```

```
. xtcorr
```

```
Estimated within-id correlation matrix R:
```

```

      c1      c2      c3      c4      c5
r1  1.0000
r2  0.0000  1.0000
r3  0.0000  0.0000  1.0000
r4  0.0000  0.0000  0.0000  1.0000
r5  0.0000  0.0000  0.0000  0.0000  1.0000

```

Any comparisons that we make between the model fits need to be made comparing models that are fit on identical data. So, for example, if you want to use a tool like `qic` to compare between correlation structures, you need to make sure that you are fitting models on identical datasets.

The final correlation structure for a GEE model that you choose for this data should depend on the importance of retaining all of the children in your analysis.

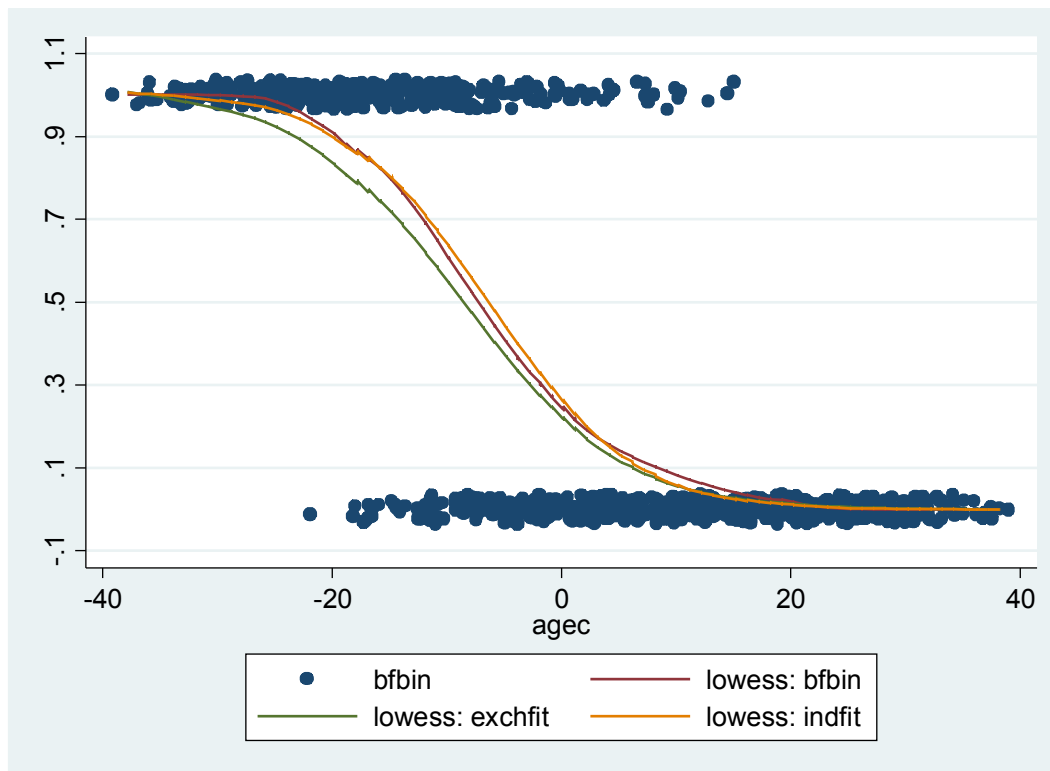
Notice that the estimates from the models fit assuming the uniform and independent correlation structures do not seem to be consistent, as they should be in the GEE models. This indicates that perhaps one of the models has still not converged, even though Stata gives no warning signs. To explore which model fits the data, we can use `xtlogit` or `gllamm` to fit the model (covered in the next lab) or do the graphical exploration that follows.

```

* get predicted values from the two models *
. quietly xtgee bfbfn male agec agemale, f(bin) link(logit) corr(exc)
nolog robust
. predict bffitexch
. label var bffitexch "exchfit"
. quietly xtgee bfbfn male agec agemale, f(bin) link(logit) corr(ind)
nolog robust
. predict bffitind
. label var bffitind "indfit"

. sort agec
* get smoothed curves of each of the sets of predictions *
. ksm bffitexch agec, lowess bw(.4) ylab(0(.2)1) lwidth(10) gen(exchsm)
. ksm bffitind agec, lowess bw(.4) ylab(0(.2)1) lwidth(10) gen(indsm)
* compare the model fits *
. twoway (scatter bfbfn agec, jitter(4)) (line bfbinsm exchsm indsm
agec, sort), ylab(-0.1(.2)1.1)

```



The independent correlation structure model (yellow) appears to fit the observed data (red) better, so we will use this as our final model on all of our 199 children.

Independence model (robust SE) results:

```

. xtgee bfbfn male agec agemale, f(bin) link(logit) corr(ind) nolog robust

GEE population-averaged model
Group variable:          id          Number of obs      =      947
Link:                   logit       Number of groups   =      199
Family:                 binomial    Obs per group: min =       1
                                           avg =      4.8

```

```

Correlation:                independent                max =          5
Wald chi2(3)                =          95.25
Scale parameter:            1                        Prob > chi2     =          0.0000
Pearson chi2(947):          739.70                Deviance        =          507.16
Dispersion (Pearson):      .7810953          Dispersion      =          .5355441

```

(Std. Err. adjusted for clustering on id)

```

-----+-----
          |                Semi-robust
          |                Coef.      Std. Err.      z      P>|z|      [95% Conf. Interval]
-----+-----
    male |      .4832121      .5502596      0.88   0.380      -.595277      1.561701
    agec |     -.2029085      .0311349     -6.52   0.000      -.2639319     -.1418852
  agemale |      .0127863      .0406788      0.31   0.753      -.0669426      .0925152
    _cons |     -1.515399      .4390092     -3.45   0.001      -2.375841     -.6549565
-----+-----

```

Get results on the odds scale

```
. xtgee, eform
```

(Std. Err. adjusted for clustering on id)

```

-----+-----
          |                Semi-robust
          |                Odds Ratio  Std. Err.      z      P>|z|      [95% Conf. Interval]
-----+-----
    male |      1.621274      .8921215      0.88   0.380      .5514098      4.766923
    agec |      .8163529      .0254171     -6.52   0.000      .7680259      .8677209
  agemale |      1.012868      .0412022      0.31   0.753      .9352489      1.09693
-----+-----

```

Population Average interpretation of coefficient on age:

Female Nepalese children of a given age have an odds of being breastfed that is 0.82 times the odds of being breastfed for female Nepalese children who are one month younger.

Test for a difference in decline of breastfeeding as children age according to gender

```
. test agemale
```

```
( 1)  agemale = 0
```

```

          chi2( 1) =      0.10
          Prob > chi2 =      0.7533

```

We have no evidence for an interaction between (male) gender and age.

Test for a gender effect in breastfeeding of Nepalese children.

```
. test male agemale
```

```
( 1)  male = 0
( 2)  agemale = 0
```

```

          chi2( 2) =      1.09
          Prob > chi2 =      0.5800

```

There is no statistically significant gender effect.

Let's next compare models that include data on 186 children:

1. GEE with independent correlation (robust SE)
2. GEE with uniform correlation (robust SE)
3. GEE with ar1 correlation (robust SE)

We first need to subset the data to just the 913 observations on 186 children that is used in the AR1 model.

```

. ** prepare to drop those individuals who were excluded in AR1 fit **
. ** save current data **
. save "nepal_temp.dta", replace
file nepal_temp.dta saved

. ** reshape to ID the dropped individuals
. reshape wide bf age agec bfbin bfbinsm agemale prob bfhat, i(id) j(visit)
(note: j = 1 2 3 4 5)

Data
-----
Number of obs.          947  ->   199
Number of variables     12  ->    43
j variable (5 values)   visit -> (dropped)
xij variables:
      bf -> bf1 bf2 ... bf5
      age -> age1 age2 ... age5
      agec -> agec1 agec2 ... agec5
      bfbin -> bfbin1 bfbin2 ... bfbin5
      bfbinsm -> bfbinsm1 bfbinsm2 ... bfbinsm5
      agemale -> agemale1 agemale2 ... agemale5
      prob -> prob1 prob2 ... prob5
      bfhat -> bfhat1 bfhat2 ... bfhat5
-----

. ** ad hoc identification of individuals dropped in AR1 model
. drop if bfbin2==.
(10 observations deleted)

. drop if bfbin4==. & bfbin5!=.
(3 observations deleted)

. reshape long
(note: j = 1 2 3 4 5)

Data
-----
Number of obs.          186  ->   930
Number of variables     43  ->    12
j variable (5 values)   ->   visit
xij variables:
      bf1 bf2 ... bf5 -> bf
      age1 age2 ... age5 -> age
      agec1 agec2 ... agec5 -> agec
      bfbin1 bfbin2 ... bfbin5 -> bfbin
      bfbinsm1 bfbinsm2 ... bfbinsm5 -> bfbinsm
      agemale1 agemale2 ... agemale5 -> agemale
      prob1 prob2 ... prob5 -> prob
      bfhat1 bfhat2 ... bfhat5 -> bfhat
-----

```

```
. xtgee bfbin male agec agemale, f(bin) link(logit) corr(ind)
GEE population-averaged model      Number of obs      =      913
Group variable:                    id      Number of groups   =      186
Link:                               logit    Obs per group: min =      3
Family:                             binomial  avg               =      4.9
Correlation:                        independent max             =      5
Scale parameter:                    1      Wald chi2(3)       =      91.15
                                           Prob > chi2        =      0.0000

Pearson chi2(913):                  703.34      Deviance           =      487.42
Dispersion (Pearson):              .7703665    Dispersion         =      .5338685
```

(Std. Err. adjusted for clustering on id)

bfbin	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
male	.4932853	.5589339	0.88	0.377	-.6022051	1.588776
agec	-.2011933	.0309748	-6.50	0.000	-.2619028	-.1404837
agemale	.0120942	.0411111	0.29	0.769	-.0684822	.0926705
_cons	-1.490251	.4392578	-3.39	0.001	-2.35118	-.6293211

```
. xtgee bfbin male agec agemale, f(bin) link(logit) corr(exch)
```

```
GEE population-averaged model      Number of obs      =      913
Group variable:                    id      Number of groups   =      186
Link:                               logit    Obs per group: min =      3
Family:                             binomial  avg               =      4.9
Correlation:                        exchangeable max             =      5
Scale parameter:                    1      Wald chi2(3)       =      106.50
                                           Prob > chi2        =      0.0000
```

(Std. Err. adjusted for clustering on id)

bfbin	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
male	-.3125047	.4385927	-0.71	0.476	-1.172131	.5471213
agec	-.1509052	.0194483	-7.76	0.000	-.1890232	-.1127871
agemale	-.033274	.0334422	-0.99	0.320	-.0988196	.0322716
_cons	-1.233789	.2822097	-4.37	0.000	-1.78691	-.6806678

```
. xtgee bfbin male agec agemale, f(bin) link(logit) corr(ar1)
```

```
GEE population-averaged model      Number of obs      =      913
Group and time vars:              id visit    Number of groups   =      186
Link:                               logit    Obs per group: min =      3
Family:                             binomial  avg               =      4.9
Correlation:                        AR(1)    max             =      5
Scale parameter:                    1      Wald chi2(3)       =      113.71
                                           Prob > chi2        =      0.0000
```

(Std. Err. adjusted for clustering on id)

bfbin	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
male	.2333108	.4723565	0.49	0.621	-.6924909	1.159113
agec	-.1780666	.023504	-7.58	0.000	-.2241337	-.1319995
agemale	.0012619	.0333198	0.04	0.970	-.0640438	.0665675

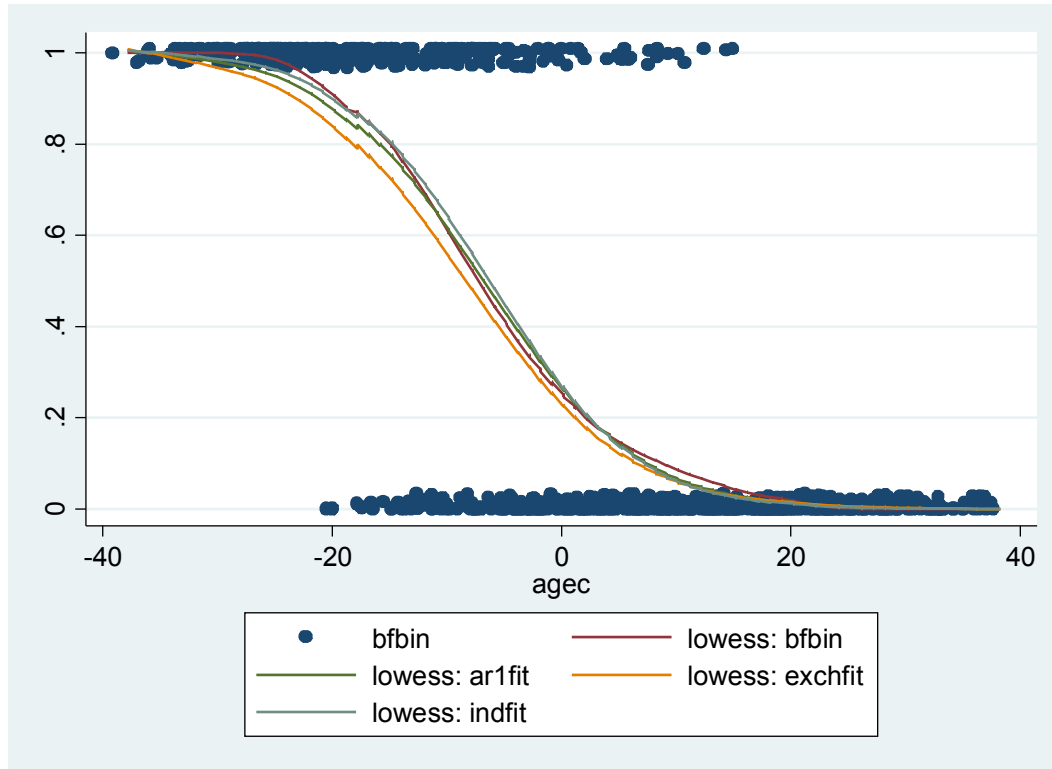
```
-----  
_cons | -1.329081 .3509273 -3.79 0.000 -2.016886 -.6412764  
-----
```

The AR1 model has estimated coefficients in the same direction as the independence model.

Let's compare the fits graphically:

Get predicted values from the three models using the robust option

```
. quietly xtgee bfbin male agec agemale, f(bin) link(logit) corr(ar1) nolog  
robust  
. predict bffitar1  
. label var bffitar1 "ar1fit"  
  
. quietly xtgee bfbin male agec agemale, f(bin) link(logit) corr(exc) nolog  
robust  
. predict bffitexch  
. label var bffitexch "exchfit"  
  
. quietly xtgee bfbin male agec agemale, f(bin) link(logit) corr(ind) nolog  
robust  
. predict bffitind  
. label var bffitind "indfit"  
  
. * need to generate a new 'data-based' smooth for the 186 children data  
. drop bfbinsm  
  
. ksm bfbin agec, lowess bw(.4) ylab(0(.2)1) lwidth(10) gen(bfbinsm)  
  
. sort age  
  
. * get smoothed curves of each of the sets of predictions *  
. ksm bffitar1 agec, lowess bw(.4) ylab(0(.2)1) lwidth(10) gen(ar1sm)  
  
. ksm bffitexch agec, lowess bw(.4) ylab(0(.2)1) lwidth(10) gen(exchsm)  
  
. ksm bffitind agec, lowess bw(.4) ylab(0(.2)1) lwidth(10) gen(indsm)  
  
. twoway (scatter bfbin agec, jitter(4)) (line bfbinsm ar1sm exchsm indsm agec,  
sort), ylab(0(.2)1)
```



AR1 and independence models appear to be better fits than the exchangeable models.

Based on the model fits for the models fit on 186 children, I would choose an AR1 model (with robust SE).

```
. xtgee bfbin male agec agemale, f(bin) link(logit) corr(ar1) nolog robust
```

```
GEE population-averaged model
Group and time vars:      id visit
Link:                     logit
Family:                   binomial
Correlation:              AR(1)
Scale parameter:         1
Number of obs             =      913
Number of groups          =      186
Obs per group: min       =         3
                       avg       =         4.9
                       max       =         5
Wald chi2(3)              =     113.71
Prob > chi2               =         0.0000
```

(Std. Err. adjusted for clustering on id)

	Coef.	Semi-robust Std. Err.	z	P> z	[95% Conf. Interval]	
male	.2333108	.4723565	0.49	0.621	-.6924909	1.159113
agec	-.1780666	.023504	-7.58	0.000	-.2241337	-.1319995
agemale	.0012619	.0333198	0.04	0.970	-.0640438	.0665675
_cons	-1.329081	.3509273	-3.79	0.000	-2.016886	-.6412764

```
. xtgee, eform
```

(Std. Err. adjusted for clustering on id)

bfbin	Odds Ratio	Semi-robust Std. Err.	z	P> z	[95% Conf. Interval]
male	1.262774	.5964794	0.49	0.621	.5003282 3.187103
agec	.8368867	.0196702	-7.58	0.000	.7992083 .8763414
agemale	1.001263	.0333619	0.04	0.970	.937964 1.068833

```
. test agemale
```

```
( 1) agemale = 0
```

```
      chi2( 1) =    0.00
      Prob > chi2 =    0.9698
```

```
. test male agemale
```

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

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

```
      chi2( 2) =    0.43
      Prob > chi2 =    0.8065
```

We come to the same conclusions about the lack of statistical significance for gender effects on breastfeeding.

You need to decide whether to choose to use AR1 (robust SE) on 186 kids or independence (robust SE) on 199 kids. We need to think about the mechanism that excludes the 13 children. If there are systematic differences between the 13 kids we exclude and the 186 kids we include, we will have to be careful about interpretations and any claims that we make about our sample being representative of a larger population.

Discussion

It seems as though our estimates of the model coefficients (especially male gender) are sensitive to the choice of the correlation structure although the estimates of the fixed effect coefficients are supposed to be consistent, regardless of the correlation structure used in GEE given that we have the correct mean structure.

Perhaps our models are not adequate for the data.

We can use xttrans to look at the transition matrix for our binary breastfeeding outcome:

```
. 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

For children who are not breastfeeding a given visit, 99.06% do not breastfeed at the next visit and 0.94% breastfeed at the next visit. For children who are breastfeeding a given visit, 13.93% do not breastfeed at the next visit and 86.07% breastfeed at the next visit.

In other words, once a child is not breastfeeding, the child usually does not start breastfeeding again. Of children who are breastfeeding, 14% will stop breastfeeding by the next visit.

Perhaps our models aren't really capturing the structure of the data. (A uniform correlation structure doesn't make much sense here.)

We'll touch on **transition models** next lab...