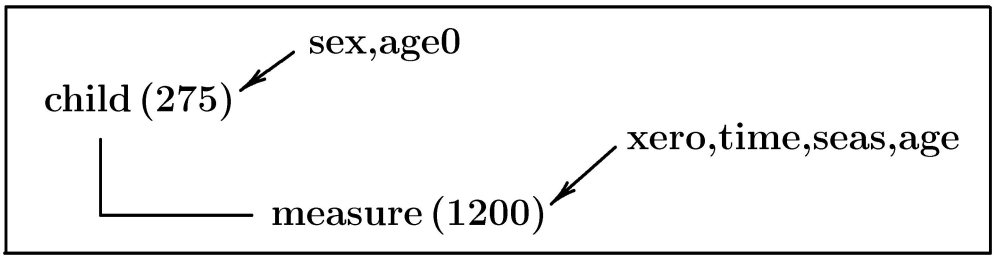
Clustered Binary Data Analysis Exercise (VER 22)
Solution

1. Identify and describe the data structure.

A: The data constitute repeated measures over time (the six consecutive quarters) on each child. We may consider this as a special case of a hierarchical data structure and display it by a diagram, including also the levels of the predictors:



Only the predictors -sex- and -age0- (child age at first measurement) are constant within each child. The numbers of units at the two levels are shown in the diagram. The replication within children is variable, with an average of 4.4 measures, a median of 4 and a range of 1-6 measures.

2. Ordinary logistic regression model-building.
It facilitates the model-building to ignore the data structure, although there is a risk that wrong decisions are made about predictors. Because accounting for the data structure will typically have the strongest impact on SEs, by making these larger, the most likely bias in the model-building is that too many predictors are carried forward to the next step – which in most cases is not a serious problem.

Besides assessing the most suitable of each predictor effect (categorical/linear/non-linear), the initial model-building step should also assess the levels of predictors (done above) and the relations between predictors. As the main focus here is not on the model-building, we list a few Stata commands and the conclusions obtained from them without showing their specific output, and then proceed to the selected simple logistic regression model.

One particular modeling decision deserves explanation. Time (i.e. quarters) could be modeled by a linear term if there was biological reason to believe that the log-odds of the respiratory disease depended linearly on time, but here one might instead anticipate a seasonal effect. Therefore time should be modeled as categorical, and we could impose an assumption that the quarters 5 and 6 should have the same disease log-odds as quarters 1 and 2 by using the 4 seasonal categories instead of the 6 time categories. However, the estimates for 6 time categories does not support such an assumption, so we abandon that idea. Furthermore, we prefer to include -age0- instead of -age- in our models, hoping to better separate the effects of time and (initial) age.

```
. tabulate time seas /* totally collinear */
. browse child time age0 age /* apparently, age increases with 3 mo. per time step */
. tabulate resp time /* no time points with all children, but most at time 1 */
. tabulate xero time /* very few cases of xero */
. logit resp xero sex i.time age0
. logit resp xero sex i.seas age0 /* much poorer fit */
. lintrend resp age0, plot(log) g(6) /* quadratic? */

. logit resp xero sex i.time c.age0##c.age0

Iteration 0:   log likelihood = -360.72665
...
Iteration 5:   log likelihood =  -326.4093
```

```
Logistic regression                               Number of obs   =       1200
                                                  LR chi2(9)      =      68.63
                                                  Prob > chi2     =     0.0000
Log likelihood =  -326.4093                       Pseudo R2       =     0.0951

------------------------------------------------------------------------------
        resp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        xero |   .7963175   .4505939     1.77   0.077    -.0868303    1.679465
         sex |   -.490786   .2242643    -2.19   0.029     -.930336   -.0512361
             |
        time |
          2  |  -1.140607   .3852212    -2.96   0.003    -1.895627   -.3855875
          3  |  -.7028209   .3584702    -1.96   0.050     -1.40541   -.0002323
          4  |  -1.394935   .4403377    -3.17   0.002    -2.257981   -.5318889
          5  |   .1835769   .2900784     0.63   0.527    -.3849663    .7521201
          6  |  -.3494412   .3190395    -1.10   0.273    -.9747472    .2758648
             |
        age0 |   .0524143   .0258527     2.03   0.043     .0017439    .1030846
             |
c.age0#c.age0 |  -.0014121   .0004489    -3.15   0.002    -.0022919   -.0005324
             |
       _cons |  -1.714057   .3569199    -4.80   0.000    -2.413607   -1.014507
------------------------------------------------------------------------------
```

We note significant effects of most predictors, including a significant quadratic term for -age0-.

3. Robust variance estimates (or standard errors).
We add the `vce(cluster child)` option to the logit command to get robust standard errors.

```
. logit resp xero sex i.time c.age0##c.age0, vce(cluster child)

Iteration 0:   log pseudolikelihood = -360.72665
...
Iteration 5:   log pseudolikelihood =  -326.4093

Logistic regression                               Number of obs   =       1200
                                                  Wald chi2(9)    =      54.92
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood =  -326.4093                 Pseudo R2       =     0.0951

                                (Std. Err. adjusted for 275 clusters in child)
------------------------------------------------------------------------------
             |               Robust
        resp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        xero |   .7963175   .4080386     1.95   0.051    -.0034234    1.596058
         sex |   -.490786   .2427091    -2.02   0.043    -.9664872   -.0150849
             |
        time |
          2  |  -1.140607   .3828259    -2.98   0.003    -1.890932   -.3902823
          3  |  -.7028209   .3577838    -1.96   0.049    -1.404064   -.0015775
          4  |  -1.394935    .400338    -3.48   0.000    -2.179583   -.6102869
          5  |   .1835769   .2869156     0.64   0.522    -.3787675    .7459212
          6  |  -.3494412   .3167455    -1.10   0.270     -.970251    .2713686
             |
        age0 |   .0524143   .0296953     1.77   0.078    -.0057874     .110616
             |
c.age0#c.age0 |  -.0014121   .0005344    -2.64   0.008    -.0024595   -.0003648
             |
       _cons |  -1.714057   .3739131    -4.58   0.000    -2.446914   -.9812011
------------------------------------------------------------------------------
```

A: Note that the listing explicitly states the use of robust clustered standard errors. Some terms in the listing have now been prefixed by “pseudo”, also to indicate that the inference is no longer (fully) model-based. Compared to the ordinary logistic model, the SEs only changed little: minor increases for -sex-, -age0- and the intercept, a small decrease for -xero- and essentially no change at all for -time-. Indeed we would have expected the SE to go up for upper-level predictors (and the intercept). The small decrease for -xero- is difficult to interpret (without

introducing more advanced concepts) and we may just accept it as is. No major changes in significance for the different effects are seen (yes, the P-value -xero- got closer to 0.05, but that is not really a major change).

4. Fixed effects modeling.
A: The idea of adding fixed effects for children to account for the data structure is seriously flawed, for two reasons. One is that this will exclude any child-level predictors from the model, and there is clearly interest in assessing effects of gender and age (represented by initial age, -age0-). The second reason is that estimating 274 fixed effects for children is likely to wreck havoc on the few coefficients we are interested in, that is, to cause substantial biases. In addition, many of the children will not show any cases of respiratory infection, and this will cause trouble for the estimation algorithm. In summary, fixed effects modeling for children is a terrible idea in this case.

5. Random effects logistic regression.
We add random effects for children, and estimate the resulting random effects logistic regression model by the adaptive Gaussian quadrature method implemented in Stata's " `meqrlogit` " command.

```
. meqrlogit resp xero sex i.time c.age0##c.age0 || child:, var

Refining starting values:
...
Iteration 2:   log likelihood = -324.42885

Performing gradient-based optimization:
...
Iteration 2:   log likelihood =  -324.4268

Mixed-effects logistic regression               Number of obs      =      1200
Group variable: child                           Number of groups   =       275

                                                Obs per group: min =         1
                                                               avg =       4.4
                                                               max =         6

Integration points =   7                        Wald chi2(9)       =     45.94
Log likelihood =  -324.4268                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
        resp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        xero |   .7038393   .4935648     1.43   0.154      -.26353    1.671209
         sex |  -.5273347   .2594732    -2.03   0.042    -1.035893   -.0187765
             |
        time |
          2  |  -1.191704   .3973365    -3.00   0.003    -1.970469    -.412939
          3  |  -.7075807   .3724488    -1.90   0.057    -1.437567    .0224056
          4  |  -1.420774   .4530175    -3.14   0.002    -2.308672   -.5328755
          5  |   .2277794   .3047732     0.75   0.455    -.3695652     .825124
          6  |  -.3426451   .3317438    -1.03   0.302    -.9928509    .3075608
             |
        age0 |   .0491582   .0291822     1.68   0.092    -.0080379    .1063542
             |
c.age0#c.age0 |  -.0013654   .0004901    -2.79   0.005     -.002326   -.0004049
             |
       _cons |  -1.854553   .4149813    -4.47   0.000    -2.667901   -1.041205
------------------------------------------------------------------------------
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
child: Identity              |
                  var(_cons) |   .5697922    .358844      .1658236    1.957883
------------------------------------------------------------------------------
LR test vs. logistic regression: chibar2(01) =     3.96 Prob>=chibar2 = 0.0232
```

A: The fit of the random effects logistic model shows a moderate improvement over the ordinary logistic

regression model, with a significant LR test (P=0.023) and an estimated between-children variance of 0.57. The clustering of responses within children (after adjustment for fixed effects) could be described as moderate, using e.g. the latent variable approximation of the ICC, calculated as: ICC = 0.57/(0.57+3.29) = 0.15. When comparing the estimates and SEs with those obtained from ordinary logistic regression (with/without SE), we need to take into account the distinction between subject-specific (SS) and population-averaged (PA) parameter values. With an estimated variance of 0.57, the SS-values from the mixed model can be converted to PA-values by dividing with the factor: sqrt(1+0.346*0.57) = 1.09, that is, a reduction of about 10%. It is seen that most of the coefficients for -sex- and -time- as well as the intercept are indeed somewhat larger in the mixed model. However, the coefficients for -xero- and -age0- are smaller in the mixed model and therefore signify substantial changes compared to the previous PA-values. We would interpret this finding as indicative of child-level confounding. In particular, one might suspect (and this can easily be verified) that the occurrence of xerophthalmia is also clustered in children. The SEs mostly follow the pattern of the estimates themselves, as we would expect, however the SE for -xero- has gone substantially up and this predictor is now clearly non-significant. It is clear that adjusting for the repeated measures impacts our estimation of -xero-. To explore this further is beyond what we can do in an introductory exercise; let us just say it has to do with -xero- itself being a time-varying predictor with clustering (see VER Sections 21.4 and 23.1.3 for some pointers).

We should also try to validate the model assumptions, but this is more difficult than for both logistic models and linear mixed models. The residuals at the lowest (measure) level are of little use due to the binary outcome (just as in ordinary logistic regression) and because of the child effects we cannot construct covariate patterns with adequate replication. The estimated random effects for children also look far from normally distributed, and this commonly happens when there is limited replication within clusters. Thus, no simple diagnostic tools are available for model checking.

6. GEE estimation.

The working correlation structures suggested by the recommendations in Lecture 11b are: exchangeable, autoregressive and unstructured. The independence structure corresponds to use of robust standard errors (discussed in 3.).

```
. xtgee resp xero sex i.time c.age0##c.age0, fam(bin) i(child) corr(exch) robust
...
Iteration 4: tolerance = 7.507e-08

GEE population-averaged model                   Number of obs      =      1200
Group variable:                       child     Number of groups   =       275
Link:                                 logit     Obs per group: min =         1
Family:                            binomial                    avg =       4.4
Correlation:                   exchangeable                    max =         6
                                                Wald chi2(9)       =     54.67
Scale parameter:                           1    Prob > chi2        =    0.0000

                                  (Std. Err. adjusted for clustering on child)
------------------------------------------------------------------------------
             |               Robust
        resp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        xero |   .7468695   .4135693     1.81   0.071    -.0637114     1.55745
         sex |  -.4894965     .24307    -2.01   0.044     -.965905   -.0130881
             |
        time |
          2  |  -1.135629       .383    -2.97   0.003    -1.886295   -.3849629
          3  |  -.6871292   .3548423    -1.94   0.053    -1.382607    .0083489
          4  |  -1.369655   .3949968    -3.47   0.001    -2.143835   -.5954759
          5  |   .1954958   .2858234     0.68   0.494    -.3647078    .7556995
          6  |  -.3369771   .3160642    -1.07   0.286    -.9564517    .2824974
             |
        age0 |   .0529615   .0295147     1.79   0.073    -.0048862    .1108092
             |
c.age0#c.age0 |  -.0014196   .0005302    -2.68   0.007    -.0024589   -.0003804
             |
       _cons |  -1.723961   .3724889    -4.63   0.000    -2.454026   -.9938962
------------------------------------------------------------------------------
```

```
. estat wcor

Estimated within-child correlation matrix R:

      |        c1         c2         c3         c4         c5         c6
------+------------------------------------------------------------------
   r1 |         1
   r2 |  .0248506          1
   r3 |  .0248506   .0248506          1
   r4 |  .0248506   .0248506   .0248506          1
   r5 |  .0248506   .0248506   .0248506   .0248506          1
   r6 |  .0248506   .0248506   .0248506   .0248506   .0248506          1

. xtset child time /* necessary preparation for GEE with time-dep structure */
       panel variable:  child (unbalanced)
        time variable:  time, 1 to 6, but with gaps
                delta:  1 unit

. xtgee resp xero sex i.time c.age0##c.age0, fam(bin) i(child) corr(ar 1) robust force
note:  some groups have fewer than 2 observations
       not possible to estimate correlations for those groups
       22 groups omitted from estimation
...
Iteration 4: tolerance = 2.648e-08

GEE population-averaged model                   Number of obs      =      1178
Group and time vars:            child time      Number of groups   =       253
Link:                                logit      Obs per group: min =         2
Family:                           binomial                     avg =       4.7
Correlation:                         AR(1)                     max =         6
                                               Wald chi2(9)       =     52.11
Scale parameter:                         1     Prob > chi2        =    0.0000

                                  (Std. Err. adjusted for clustering on child)
------------------------------------------------------------------------------
             |               Robust
        resp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        xero |   .7475698   .4154792     1.80   0.072    -.0667545    1.561894
         sex |  -.5038004   .2459155    -2.05   0.040    -.9857858   -.0218149
             |
        time |
          2  |  -1.092161   .3832552    -2.85   0.004    -1.843328   -.3409949
          3  |  -.6732309   .3601884    -1.87   0.062    -1.379187    .0327255
          4  |  -1.372688   .4030994    -3.41   0.001    -2.162748   -.5826272
          5  |   .1890115   .2920394     0.65   0.517    -.3833752    .7613982
          6  |  -.3357932   .3206422    -1.05   0.295    -.9642404    .2926539
             |
        age0 |     .05575   .0299245     1.86   0.062    -.0029009    .1144009
             |
c.age0#c.age0 |  -.0014531   .0005402    -2.69   0.007    -.0025118   -.0003945
             |
       _cons |  -1.766946   .3764598    -4.69   0.000    -2.504794   -1.029099
------------------------------------------------------------------------------

. estat wcor

Estimated within-child correlation matrix R:

      |        c1         c2         c3         c4         c5         c6
------+------------------------------------------------------------------
   r1 |         1
   r2 |  .0293358          1
   r3 |  .0008606   .0293358          1
   r4 |  .0000252   .0008606   .0293358          1
   r5 |  7.41e-07   .0000252   .0008606   .0293358          1
   r6 |  2.17e-08   7.41e-07   .0000252   .0008606   .0293358          1

. xtgee resp xero sex i.time c.age0##c.age0, fam(bin) i(child) corr(uns) robust

Iteration 1: tolerance = .02133025
...
Iteration 5: tolerance = 3.319e-07
```

```
GEE population-averaged model                   Number of obs      =      1200
Group and time vars:            child time      Number of groups   =       275
Link:                                logit      Obs per group: min =         1
Family:                            binomial                     avg =       4.4
Correlation:                  unstructured                      max =         6
                                                Wald chi2(9)       =     57.37
Scale parameter:                         1      Prob > chi2        =    0.0000

                                  (Std. Err. adjusted for clustering on child)
------------------------------------------------------------------------------
             |               Robust
        resp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        xero |   .7812751   .4018158     1.94   0.052    -.0062694     1.56882
         sex |  -.4707663   .2427171    -1.94   0.052     -.946483    .0049504
             |
        time |
          2  |  -1.151599   .3846329    -2.99   0.003    -1.905466   -.3977328
          3  |  -.7095029   .3570374    -1.99   0.047    -1.409283   -.0097224
          4  |  -1.395854   .3985818    -3.50   0.000     -2.17706   -.6146478
          5  |   .2091062   .2837769     0.74   0.461    -.3470863    .7652987
          6  |  -.3567317   .3171539    -1.12   0.261    -.9783419    .2648785
             |
        age0 |   .0532271   .0290305     1.83   0.067    -.0036716    .1101257
             |
c.age0#c.age0 |  -.0014262   .0005221    -2.73   0.006    -.0024496   -.0004028
             |
        _cons |  -1.729522   .3679945    -4.70   0.000    -2.450778   -1.008266
------------------------------------------------------------------------------

. estat wcor

Estimated within-child correlation matrix R:

      |        c1         c2         c3         c4         c5         c6
------+------------------------------------------------------------------
   r1 |         1
   r2 |  .0181911          1
   r3 |  .0074442   .0463014          1
   r4 |  .0963457  -.0229863  -.0066756          1
   r5 |   .009469   .1340138   .0820219  -.0333444          1
   r6 | -.0039283   .0474818  -.0536404  -.0083322   .0491443          1
```

A: The estimates, standard errors and P-values are quite similar across the different GEE estimations. Both -sex- and -xero- have Wald tests with P-values just around 0.05. The estimated working correlation matrices differ quite a bit between different GEE versions, but as all correlations are small, these differences have no major impact on the inference for fixed effects. With the very large number of subjects, this serves to illustrate the asymptotic unbiasedness of GEE estimates, regardless of working correlation structure. It is surprising that the exchangeable working correlation is estimated as low as 0.025, while the random effects model produced an approximate ICC of 0.15. The analysis in the textbook Diggle et al. (2002) (referred to in the problem) used the ALR version of GEE which gives a better estimate for the within-subject association.

In summary, the random effects model estimate for -xero- differed somewhat from those of the other approaches, but otherwise there was good agreement between the different approaches.