Module 2

Introduction to Longitudinal Data Analysis

Case Studies

Colleen Sitlani, PhD
Cardiovascular Health Research Unit
Department of Medicine
University of Washington

SISCER
July 22, 2019

Overview

Case Study: Longitudinal Depression Scores

Case Study: Indonesia Children's Health Study

Depression Study: Motivation and Design

- Gregoire et al (1996) published the results of an efficacy study on estrogen patches in treating postnatal depression
- 61 women with major depression, which began within 3 months of childbirth and persisted for up to 18 months postnatally, participated in a double-blind, placebo-controlled study
- Women were randomly assigned to active treatment (n=34) or placebo (n=27)
- Participants attended clinics monthly and at each visit self-ratings of depressive symptoms on the Edinburgh postnatal depression scale (EPDS) were measured
- EPDS is a standardized, validated, self-rating scale consisting of 10 items, each rated on a 4-point scale of 0–3
- **Goal**: Investigate the antidepressant efficacy of treatment with estrogen over time

Depression Study: Data

- Depression scores are assessed across $m = 7$ months for the $n = 61$ subjects in the study
- Depression scores for visit j are the longitudinal components measured on subject i

	subj	group	dep0	dep1	dep2	dep3	dep4	dep5	dep6
1.	1	placebo	18	17	18	15	17	14	15
2.	2	placebo	27	26	23	18	17	12	10
3.	3	placebo	16	17	14	.	.	.	.
4.	4	placebo	17	14	23	17	13	12	12
5.	5	placebo	15	12	10	8	4	5	5
6.	6	placebo	20	19	11.54	9	8	6.82	5.05
7.	7	placebo	16	13	13	9	7	8	7
8.	8	placebo	28	26	27	.	.	.	.
9.	9	placebo	28	26	24	19	13.94	11	9
10.	10	placebo	25	9	12	15	12	13	20

- 'Wide' form: A row for each subject
- Note that there are some missing data due to drop-out

Depression Study Questions: EDA

1. Summarize the depression scores by visit and treatment group.
2. Examine within-person correlations among depression scores, graphically and numerically.
3. Graph depression scores over time, by treatment group. Include a lowess line (smoother) for each group to summarize trends.
4. Plot individual trajectories by treatment group.

Depression Study Questions: Regression Analyses

5. Consider collapsing the longitudinal series for each subject into a summary statistic between the baseline and sixth depression scores. Use methods for independent data to evaluate the association between change in depression scores and estrogen treatment.
6. Reshape the data into long form and evaluate longitudinal associations between depression scores and treatment using GEE.
   - Use visit as a linear variable.
   - Use visit as a categorical variable.
   - Evaluate whether the treatment effect varies over time.

Reshape the Data

Recall what the data look like in wide form

	subj	group	dep0	dep1	dep2	dep3	dep4	dep5	dep6
1.	1	placebo	18	17	18	15	17	14	15
2.	2	placebo	27	26	23	18	17	12	10
3.	3	placebo	16	17	14	.	.	.	.
4.	4	placebo	17	14	23	17	13	12	12
5.	5	placebo	15	12	10	8	4	5	5

For some analyses, reshape the data from wide form to long form

```
. reshape long dep, i(subj) j(visit)
(note: j = 0 1 2 3 4 5 6)

Data                               wide   ->   long
-----------------------------------------------------
Number of obs.                       61   ->     427
Number of variables                   9   ->       4
j variable (7 values)                     ->   visit
xij variables:
                     dep0 dep1 ... dep6   ->   dep
-----------------------------------------------------
```

Reshape the Data

'Long' form: A row for each observation

	subj	visit	group	dep
1.	1	0	placebo	18
2.	1	1	placebo	17
3.	1	2	placebo	18
4.	1	3	placebo	15
5.	1	4	placebo	17
6.	1	5	placebo	14
7.	1	6	placebo	15
8.	2	0	placebo	27
9.	2	1	placebo	26
10.	2	2	placebo	23

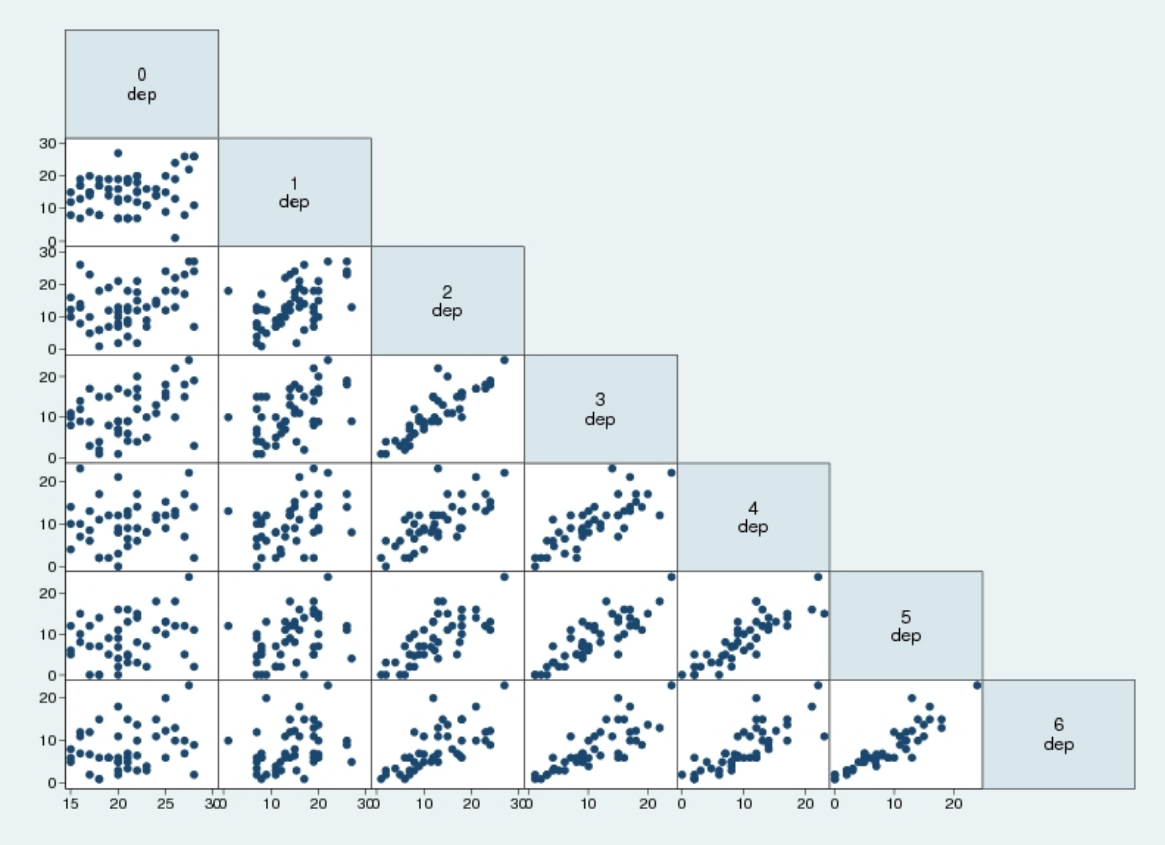
Summarize by group and visit

```
. sort group
. by group: summarize dep0 dep1 dep2 dep3 dep4 dep5 dep6
-------------------------------------------------------------------------------
-> group = placebo
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        dep0 |        27    20.77778    3.954874        15         28
        dep1 |        27    16.48148    5.279644         7         26
        dep2 |        22    15.88818    6.124177         4         27
        dep3 |        17    14.12882    4.974648      4.19         22
        dep4 |        17    12.27471    5.848791         2         23
        dep5 |        17    11.40294    4.438702      3.03         18
        dep6 |        17    10.89588     4.68157      3.45         20
-------------------------------------------------------------------------------
-> group = estrogen
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        dep0 |        34    21.24882    3.574432        15         28
        dep1 |        34    13.36794    5.556373         1         27
        dep2 |        31    11.73677    6.575079         1         27
        dep3 |        29    9.134138    5.475564         1         24
        dep4 |        28    8.827857    4.666653         0         22
        dep5 |        28    7.309286    5.740988         0         24
        dep6 |        28    6.590714    4.730158         1         23
-------------------------------------------------------------------------------
```

- **Note**: There are fewer observations observed over time

Correlation

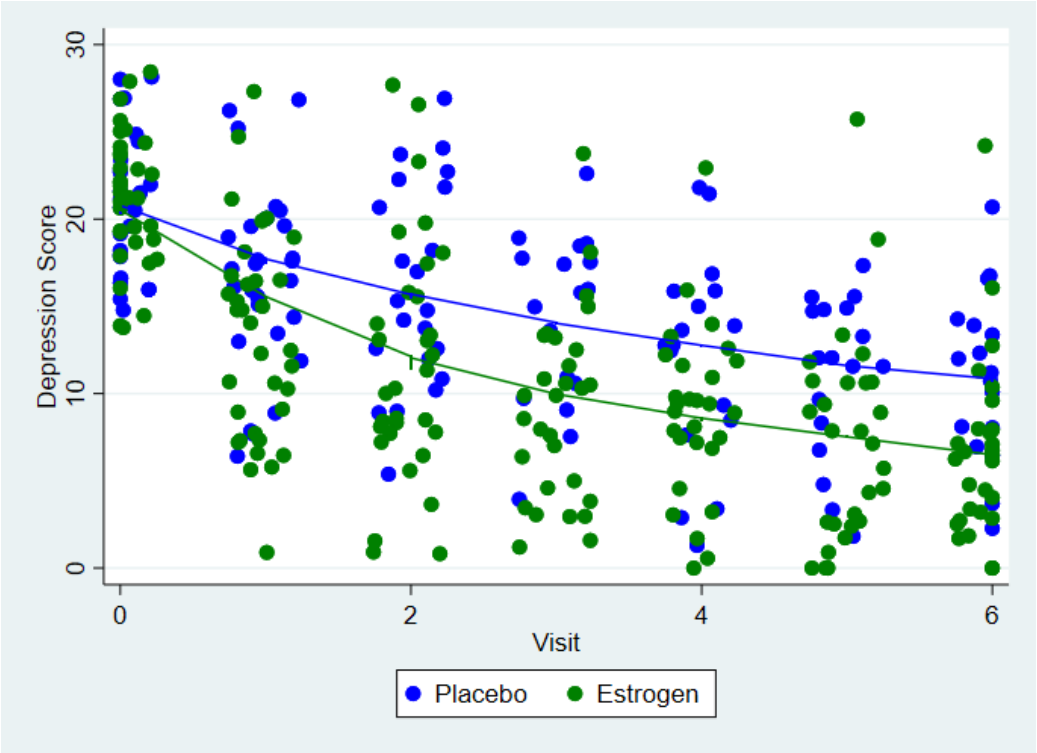
```
. graph matrix dep0 dep1 dep2 dep3 dep4 dep5 dep6, half
```



- All correlations are positive
- Strong correlation between adjacent visits

Depression scores over time

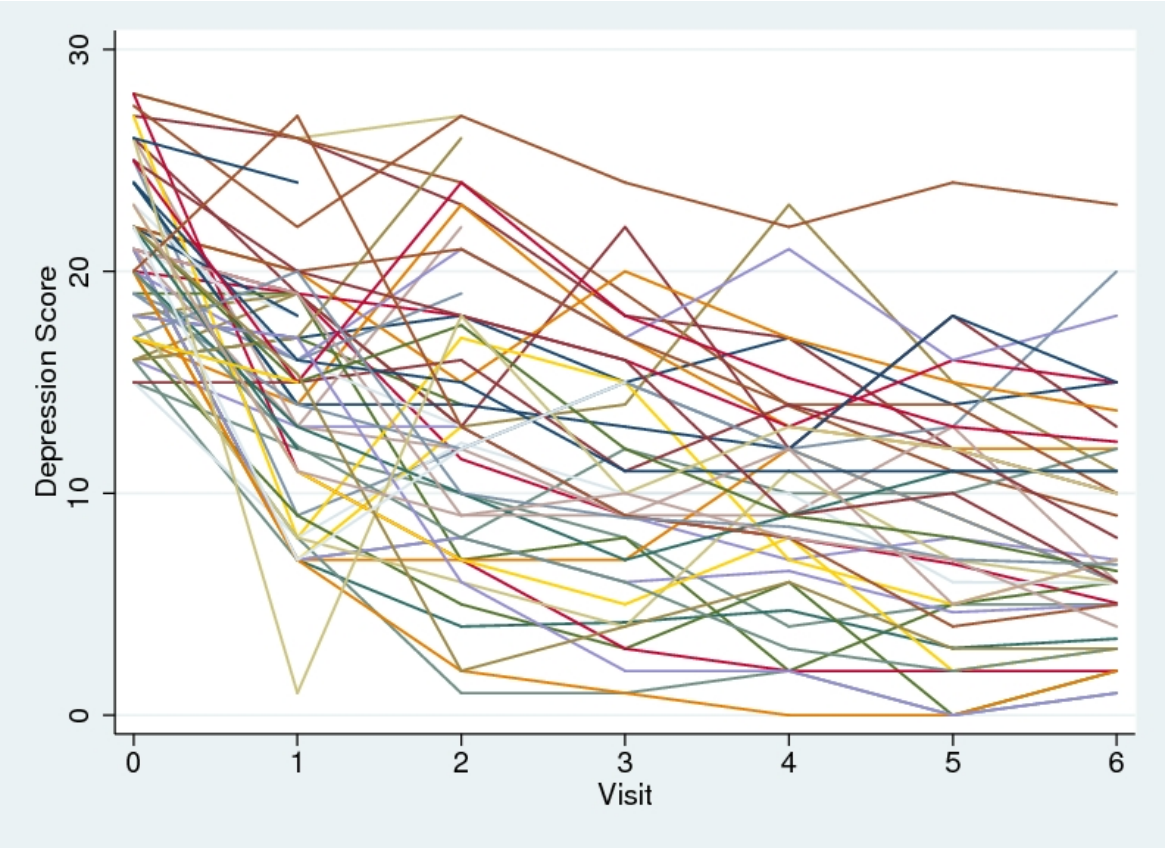
```
. separate dep, by(group)
. graph twoway (scatter dep0 visit, jitter(10) mcolor(blue))
    (scatter dep1 visit, jitter(10) mcolor(green)) ///
    (lowess dep0 visit, lcolor(blue)) (lowess dep1 visit, lcolor(green))
```



- For each treatment arm, mean depression scores decrease over time

Individual Trajectories

```
. xtline dep, i(subj) t(visit) overlay legend(off)
    xlab(0(1)6) xtitle("Visit") ytitle("Depression Score")
```



- Reveals the complexity of individual trajectories
- Note that several patients drop out after the second visit

Simple difference

```
. gen diff=dep6-dep0
(16 missing values generated)

. ttest diff, by(group) unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 placebo |      17   -9.633529    1.321784    5.449855   -12.43559   -6.831472
estrogen |      28   -14.71143    .8682517    4.594356   -16.49293   -12.92992
---------+--------------------------------------------------------------------
combined |      45   -12.79311    .8158414     5.47283   -14.43733   -11.14889
---------+--------------------------------------------------------------------
    diff |            5.077899    1.581447                1.845991    8.309808
------------------------------------------------------------------------------
    diff = mean(placebo) - mean(estrogen)                         t =   3.2109
Ho: diff = 0                     Satterthwaite's degrees of freedom =  29.5287

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9984         Pr(|T| > |t|) = 0.0032          Pr(T > t) = 0.0016
```

- Clear decreases over time; larger decreases among estrogen group
- Limited to those with complete measurement series

Generalized Estimating Equations (GEE)

- A special feature of longitudinal data is that the $m = 7$ observations that are nested within the $n = 61$ subjects are ordered in time
- We can consider *marginal models* to model the within-subject dependence by allowing us to specify the covariance structure across the nested observations
- Parameters describing the covariance must be estimated along with traditional regression coefficients
- A variety of options are available to describe the covariance
- Some covariance patterns require more information, i.e., require more parameters to be estimated than others
- Recall, we identify the data as a 'panel' data set using the `xtset` command in Stata

Assumptions

To account for the repeated measures we can use generalized estimating equations which include all of the data over the time points in a marginal model for the mean response and account for the longitudinal correlation

$$g\left(E[Y_{ij} \mid x_{ij}]\right) = x_{ij}\beta \quad \text{and} \quad Corr[Y_{ij}, Y_{ij'}] = \rho(\alpha)$$

Assumptions

- Observations are independent across subjects
- Observations may be correlated within subjects
- Missing data are missing completely at random

Statistical Model

- Using the GEE framework, we consider the 'cross-sectional' model where we are interested in the average treatment effect over time

$$E[Y_{ij} \mid x_{ij}] = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2}$$

where

  - Y_{ij}: continuous depression score (`dep`)
  - x_{ij1}: continuous variable for visit (`visit`)
  - x_{ij2}: binary treatment group with 1=treatment, 0=placebo (`group`)

- For the continuous outcome, we use an identity link, `link(iden)`, in the Gaussian family, `fam(gaus)`; these are the default
- In Stata, `xtgee` allows us to specify various working covariance structures through the `corr` option. The command `estat wcorr` allows us to view the working correlation matrix

Common Correlation Structures

- **Independence**: Observations are assumed to be independent
  - For correlation between any two observations on the same subject we assume that $\text{Corr}[Y_{ij}, Y_{ij'}] = 0$
  - It is unlikely that for any subject, depression scores are independent from one visit to the next
- **Exchangeable**: Correlations are assumed to be constant between any two observations on the same subject; $\text{Corr}[Y_{ij}, Y_{ij'}] = \alpha$
- **AR(1)**: Correlation is assumed to decay as a function of time or distance between observations; $\text{Corr}[Y_{ij}, Y_{ij'}] = \alpha^{|j-j'|}$
  - Likely to be appropriate in cases where there are a reasonable number of repeated measurements over time
  - Given that our data are measured over time, using the AR(1) correlation may help increase efficiency of SE estimation
- **Unstructured**: No relationship is imposed on dependence over time or within subjects; $\text{Corr}[Y_{ij}, Y_{ij'}] = \alpha_{jj'}$

⋆ Robust variance estimator protects against incorrect choice

GEE-independence

```
. xtgee dep visit i.group, corr(ind) robust

Iteration 1: tolerance = 9.496e-16

GEE population-averaged model                   Number of obs     =        356
Group variable:                        subj     Number of groups  =         61
Link:                                identity   Obs per group:
Family:                              Gaussian                 min =          2
Correlation:                      independent                 avg =        5.8
                                                              max =          7
                                                Wald chi2(2)      =     188.72
Scale parameter:                     29.02175   Prob > chi2       =     0.0000

Pearson chi2(356):                   10331.74   Deviance          =   10331.74
Dispersion (Pearson):                29.02175   Dispersion        =   29.02175

                                  (Std. Err. adjusted for clustering on subj)
------------------------------------------------------------------------------
             |               Robust
         dep |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       visit |  -1.921912   .1413007   -13.60   0.000    -2.198857   -1.644968
             |
       group |
   estrogen  |  -3.208912    1.08604    -2.95   0.003    -5.337511   -1.080313
       _cons |   20.19473   .8278936    24.39   0.000     18.57209    21.81737
------------------------------------------------------------------------------
```

GEE-AR(1)

```
. xtgee dep visit i.group, corr(ar1) robust

Iteration 1: tolerance = .14319978
...
Iteration 7: tolerance = 6.710e-07

GEE population-averaged model                   Number of obs     =        356
Group and time vars:            subj visit      Number of groups  =         61
Link:                             identity      Obs per group:
Family:                           Gaussian                    min =          2
Correlation:                         AR(1)                    avg =        5.8
                                                              max =          7
                                                Wald chi2(2)      =     255.61
Scale parameter:                   29.8609      Prob > chi2       =     0.0000

                                  (Std. Err. adjusted for clustering on subj)
------------------------------------------------------------------------------
             |               Robust
         dep |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       visit |  -2.073222   .1300662   -15.94   0.000    -2.328147   -1.818297
             |
       group |
   estrogen  |  -2.529295   .9610062    -2.63   0.008    -4.412832   -.6457574
       _cons |   21.01002   .7325074    28.68   0.000     19.57433    22.44571
------------------------------------------------------------------------------
```

Working correlation structure

Examine the correlation structure estimated by the model

```
. estat wcorr

Estimated within-subj correlation matrix R:
      |        c1         c2         c3         c4         c5         c6         c7
------+-----------------------------------------------------------------------------
   r1 |         1
   r2 |  .6447567          1
   r3 |  .4157113   .6447567          1
   r4 |  .2680326   .4157113   .6447567          1
   r5 |  .1728158   .2680326   .4157113   .6447567          1
   r6 |  .1114242   .1728158   .2680326   .4157113   .6447567          1
   r7 |  .0718415   .1114242   .1728158   .2680326   .4157113   .6447567          1
```

Compare with simple pairwise correlations

```
. corr dep0 dep1 dep2 dep3 dep4 dep5 dep6
(obs=45)

             |     dep0     dep1     dep2     dep3     dep4     dep5     dep6
-------------+---------------------------------------------------------------
        dep0 |   1.0000
        dep1 |   0.1922   1.0000
        dep2 |   0.3904   0.4982   1.0000
        dep3 |   0.3958   0.5258   0.8672   1.0000
        dep4 |   0.1658   0.3933   0.7357   0.7831   1.0000
        dep5 |   0.2848   0.3674   0.7500   0.8520   0.8449   1.0000
        dep6 |   0.2688   0.2795   0.6900   0.7967   0.7894   0.9014   1.0000
```

Modeling time

- Valid inference from GEE requires that the mean model is correct
- We have two covariates: treatment group is binary, time is ?
- Instead of a continuous variable (or 'grouped linear' term) for time, consider a categorical variable

$$\begin{aligned} \mathrm{E}[Y_{ij}|x_{ij}] = \beta_0 + \beta_2 x_{ij2} + \beta_3 x_{ij3} + \beta_4 x_{ij4} \\ + \beta_5 x_{ij5} + \beta_6 x_{ij6} + \beta_7 x_{ij7} + \beta_8 x_{ij8} \end{aligned}$$

with, in addition to x_{ij2} representing the treatment variable (`group`)

- x_{ij3}: dummy variable for visit 1 compared to visit 0
- x_{ij4}: dummy variable for visit 2 compared to visit 0

  ⋮

- x_{ij8}: dummy variable for visit 6 compared to visit 0

GEE-AR(1), categorical time

```
. xtgee dep i.group i.visit, corr(ar1) robust

Iteration 1: tolerance = .13810114
...
Iteration 5: tolerance = 7.593e-08

GEE population-averaged model                   Number of obs     =        356
Group and time vars:                 subj visit Number of groups  =         61
Link:                                  identity Obs per group:
Family:                                Gaussian                min =          2
Correlation:                              AR(1)                avg =        5.8
                                                               max =          7
                                                Wald chi2(7)      =     288.60
Scale parameter:                        26.7531 Prob > chi2       =     0.0000

                                  (Std. Err. adjusted for clustering on subj)
------------------------------------------------------------------------------
             |               Robust
         dep |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |
   estrogen  |  -2.593467   .9610867    -2.70   0.007    -4.477163    -.709772
             |
       visit |
          1  |  -6.294262   .7775699    -8.09   0.000    -7.818271   -4.770253
          2  |  -7.341596   .8475509    -8.66   0.000    -9.002766   -5.680427
          3  |  -9.258931   .7719962   -11.99   0.000    -10.77202   -7.745847
          4  |  -10.25842   .8352919   -12.28   0.000    -11.89557   -8.621282
          5  |  -11.69253    .807447   -14.48   0.000     -13.2751   -10.10997
          6  |  -12.43824   .7614791   -16.33   0.000    -13.93071   -10.94577
             |
       _cons |   22.48587   .7687195    29.25   0.000      20.9792    23.99253
------------------------------------------------------------------------------
```

Modeling time

- Strong evidence that depression scores vary over time

```
. testparm i.visit

 ( 1)  1.visit = 0
 ( 2)  2.visit = 0
 ( 3)  3.visit = 0
 ( 4)  4.visit = 0
 ( 5)  5.visit = 0
 ( 6)  6.visit = 0

           chi2(  6) =  287.46
         Prob > chi2 =    0.0000
```

- In the model with continuous visit, the difference in mean score between groups was -2.53 and it was highly significant ($p = 0.008$)
- When considering categorical visit, the difference in mean score between groups was -2.59 and it was highly significant ($p = 0.007$)
- Noting that the estimated treatment effect is the same in both models, we opt for the parsimony of the model with continuous visit

Model with Interaction

Consider a model that allows the treatment effect to depend on time

- The model of interest becomes

$$E[Y_{ij} \mid x_{ij}] = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 (x_{ij1} \times x_{ij2})$$

  where Y_{ij} is the continuous depression score, x_{ij1} is a continuous variable for visit and x_{ij2} is the treatment variable

- Model includes the main effects and the interaction term
- For subjects in the placebo group ($x_{ij2} = 0$), the model is

$$E[Y_{ij} \mid x_{ij}] = \beta_0 + \beta_1 x_{ij1}$$

- For subjects in the estrogen group ($x_{ij2} = 1$), the model is

$$E[Y_{ij} \mid x_{ij}] = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_{ij1}$$

- Now we can compare whether the mean change in depression score over time differs between treatment groups ('longitudinal' model)

GEE-AR(1), continuous time, interaction

```
. xtgee dep c.visit##i.group, corr(ar1) robust


Iteration 1: tolerance = .34080358
...
Iteration 8: tolerance = 2.747e-07

GEE population-averaged model                   Number of obs     =        356
Group and time vars:                 subj visit Number of groups  =         61
Link:                                  identity Obs per group:
Family:                                Gaussian                min =          2
Correlation:                              AR(1)                avg =        5.8
                                                               max =          7
                                                Wald chi2(3)      =     325.29
Scale parameter:                       29.59602 Prob > chi2       =     0.0000

                                  (Std. Err. adjusted for clustering on subj)
------------------------------------------------------------------------------
             |               Robust
         dep |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       visit |  -1.645136   .2032329    -8.09   0.000    -2.043465   -1.246807
             |
       group |
    estrogen |   -.668246   .9514551    -0.70   0.482    -2.533064    1.196572
             |
group#c.visit |
    estrogen |  -.7209406    .250909    -2.87   0.004    -1.212713   -.2291681
             |
       _cons |    19.9757   .7700831    25.94   0.000     18.46636    21.48503
------------------------------------------------------------------------------
```

Interpretation of model with interaction

- Estimate the change over time for the estrogen group by adding the coefficients for the visit variable and the interaction term

```
. lincom visit + 1.group#c.visit

 ( 1)  visit + 1.group#c.visit = 0

------------------------------------------------------------------------------
         dep |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -2.366076   .1471451   -16.08   0.000    -2.654475   -2.077677
------------------------------------------------------------------------------
```

- For a population of women on placebo treatment, mean depression score decreases by approximately 1.65 points for each additional visit, 95% CI: (-2.04, -1.25)
- For a population of women on estrogen treatment, mean depression score decreases by approximately 2.37 points for each additional visit, 95% CI: (-2.65, -2.08)
- Strong evidence that these associations are different ($p = 0.004$)

Conclusion

- GEE is specified by a mean model and a correlation model.
  - We created a linear model for the average depression score and modeled the longitudinal correlation using the AR(1) structure
- GEE requires that the mean model is correctly specified
  - We explored different options for modeling temporal trends
- GEE provides valid estimates and standard errors for the regression parameters even under misspecification of the correlation structure, but efficiency gains are possible if the correlation model is correct
  - We chose AR(1) with the robust option
- Model with a group-by-time interaction term facilitated estimation of changes over time within groups and between-group comparisons in temporal trends
  - Contrasted this with a cross-sectional model that compared the mean depression score between groups over all times

Overview

Case Study: Longitudinal Depression Scores

Case Study: Indonesia Children's Health Study

Indonesia Children's Health Study (ICHS)

- Determine the effects of vitamin A deficiency in preschool children
- $n = 275$ children examined for respiratory infection at up to 6 visits
- Xeropthalmia is an ocular manifestation of vitamin A deficiency
- **Goal**: Evaluate association between vitamin A deficiency and risk of respiratory infection

		Age (years)							
Xeropthalmia	Infection	0	1	2	3	4	5	6	7
No	No	77	229	154	196	176	143	65	5
No	Yes	8	30	30	15	9	7	1	0
Yes	No	0	1	9	10	15	8	4	1
Yes	Yes	0	0	4	3	0	0	0	0

ICHS: Data

```
. list id age time infection xerop gender hfora cost sint
```

	id	age	time	infect~n	xerop	gender	hfora	cost	sint
1.	121013	31	1	0	0	0	-3	-1	0
2.	121013	34	2	0	0	0	-3	0	-1
3.	121013	37	3	0	0	0	-2	1	0
4.	121013	40	4	0	0	0	-2	0	1
5.	121013	43	5	1	0	0	-2	-1	0
6.	121013	46	6	0	0	0	-3	0	-1
7.	121113	-9	1	0	0	1	2	-1	0
8.	121113	-6	2	0	0	1	0	0	-1
9.	121113	-3	3	0	0	1	-1	1	0
10.	121113	0	4	0	0	1	-2	0	1
11.	121113	3	5	1	0	1	-3	-1	0
12.	121113	6	6	0	0	1	-3	0	-1
13.	121114	-26	1	0	0	0	8	-1	0
14.	121114	-23	2	0	0	0	5	0	-1
15.	121114	-20	3	0	0	0	3	1	0
16.	121114	-17	4	1	0	0	0	0	1
17.	121114	-14	5	1	0	0	0	-1	0
18.	121114	-11	6	0	0	0	0	0	-1

Multiple records per person, with age in months, centered at 36 months, and time indicating visit number

ICHS Questions: EDA

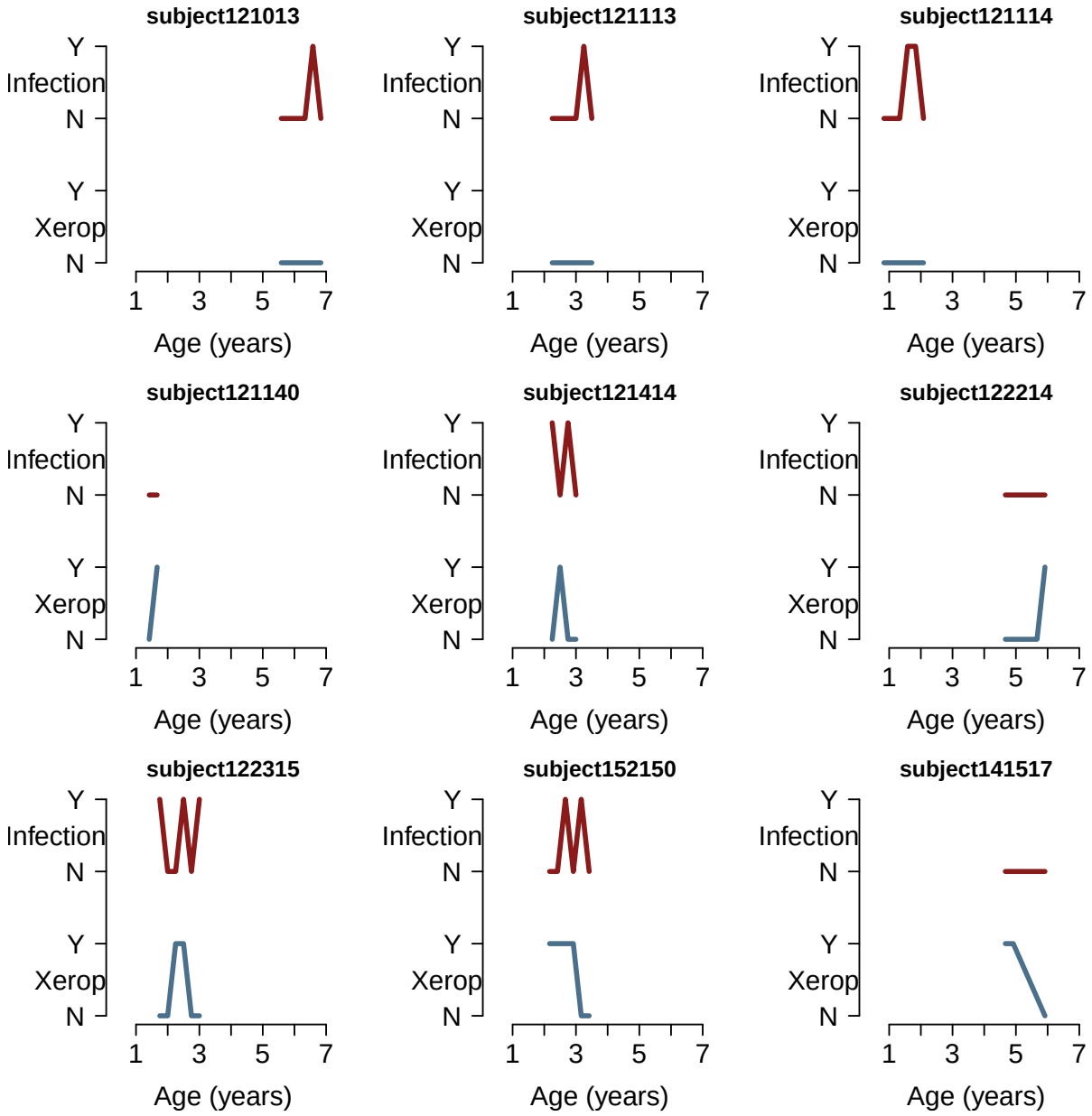
1. Plot vitamin A deficiency and infection status, by age, for a sample of individuals.
2. Plot percent with respiratory infection versus age, by presence or absence of vitamin A deficiency.
3. Explore correlation structure by visit number, and calculate percent with respiratory infection at each visit.

ICHS Questions: Regression Analyses

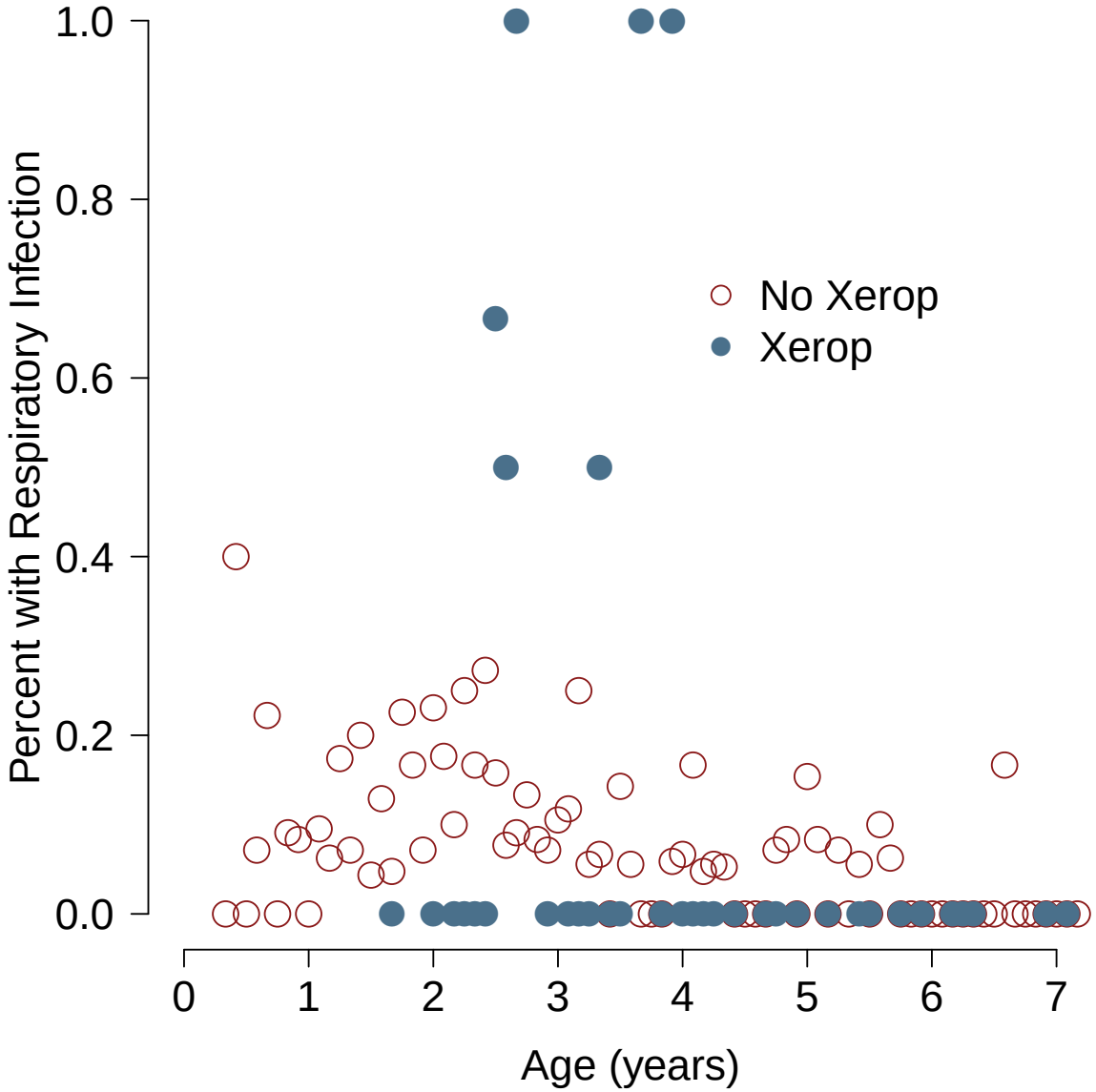
4. Evaluate the association between respiratory infection and vitamin A deficiency using an ordinary logistic regression model.
5. Use GEE to estimate the population-averaged odds ratio for respiratory infection, comparing those with vitamin A deficiency to those without, given equivalent values of other covariates. Explore multiple specifications of working correlation.

6. Use GLMM to estimate the conditional odds ratio for respiratory infection, comparing a typical individual with vitamin A deficiency to a typical individual without, given equivalent values of other covariates. Estimate the variability in the probability of respiratory infection across individuals.

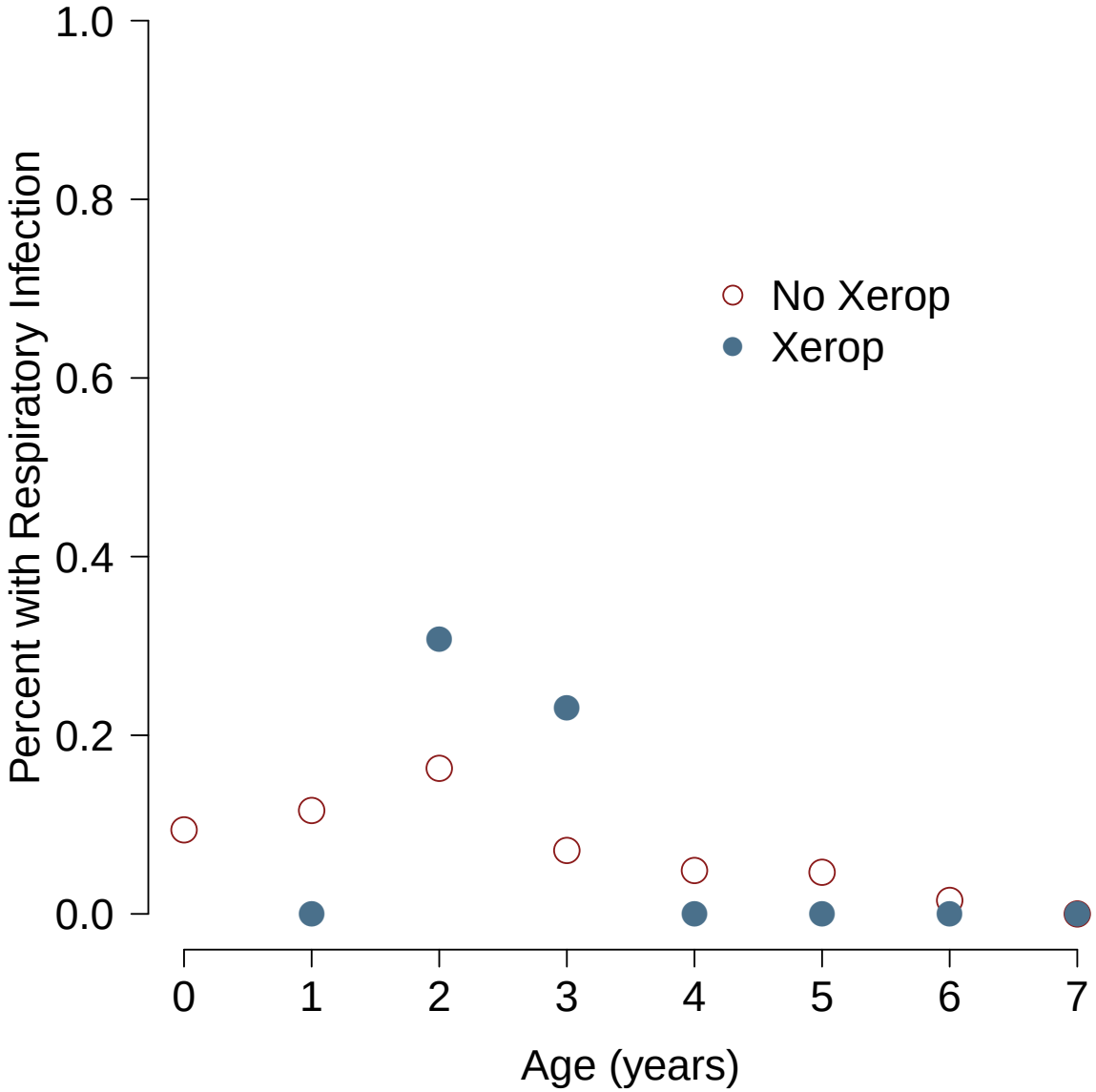
Individual trajectories



Monthly averages



Yearly averages



Logistic regression model

```
> summary(glm(infection ~ xerop + age + gender + hfora + cost + sint,
                data=ichs, family="binomial"))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.42134    0.15920  -15.21  < 2e-16 ***
xerop        0.73148    0.43591    1.68  0.09334 .
age         -0.03188    0.00634   -5.03  4.9e-07 ***
gender      -0.39364    0.21965   -1.79  0.07311 .
hfora       -0.04944    0.02012   -2.46  0.01401 *
cost        -0.58029    0.16722   -3.47  0.00052 ***
sint        -0.16536    0.16851   -0.98  0.32645
```

- $\exp(\beta_1) = 2.08$
- 95%CI = (0.88, 4.88)
- Does not take into account within-person correlation

GEE motivation

Do vitamin A deficient children have an increased risk of infection?

$$\begin{aligned} \mu_{ij} &= \mathrm{E}[Y_{ij} \mid x_{ij}] \\ &= \mathrm{P}[Y_{ij} = 1 \mid x_{ij}] \end{aligned}$$

$$\begin{aligned} \operatorname{logit} \mu_{ij} &= \log \frac{\mu_{ij}}{1 - \mu_{ij}} \\ &= \beta_0 + \beta_1 \, \mathrm{Xerophthalmia}_{ij} + \cdots \\ &\approx \log \frac{\mathrm{P}[Y_{ij} = 1 \mid x_{ij}]}{\mathrm{P}[Y_{ij} = 0 \mid x_{ij}]} \end{aligned}$$

- $\exp \beta_1$ represents the ratio of the expected odds of respiratory infection among a population of vitamin A deficient children to that for a population of children replete with vitamin A of the same age, gender, ...
- $\exp \beta_1$ is therefore a **population-averaged** parameter
- Respiratory infection is rare so odds ratio approximates relative risk

Correlations

- Use visit time (not age) to obtain a correlation matrix with $n = 146$–229 observations per cell

	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6
Time 1	1					
Time 2	0.06	1				
Time 3	0.07	0.11	1			
Time 4	0.24	-0.03	0.06	1		
Time 5	0.07	0.26	0.19	-0.01	1	
Time 6	0.05	0.12	-0.07	0.06	0.10	1
Infection	13 %	5 %	7 %	4 %	15 %	9 %

Comments on covariance structure

- For a binary outcome, variance depends on mean

$$\text{Var}(Y_{ij}) = \text{E}[Y_{ij}](1 - \text{E}[Y_{ij}])$$

- Correlation also depends (in a somewhat complicated way) on pairwise means
- **NB**
  - With respect to age, data are neither balanced nor complete
  - Even if our analysis will be a function of age, examination of covariance and correlation matrices with respect to visit time is useful
  - Dependence of correlation on pairwise means motivates alternate methods that model odds ratios instead of correlations

Covariance structure

- Odds ratios measure the association between two binary variables
- Here, binary outcomes at two different visit times

	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6
Time 1	∞					
Time 2	1.93	∞				
Time 3	2.10	4.62	∞			
Time 4	8.60	0	2.38	∞		
Time 5	1.76	11.9	4.68	0.92	∞	
Time 6	1.63	3.73	0	2.18	2.14	∞

Covariance structure

- Variance model

$$\text{Var}[Y_{ij} \mid x_{ij}] = \mu_{ij}(1 - \mu_{ij})$$

- Consider various specifications for the 'working' correlation structure
  - Independence
  - Exchangeable
  - Auto-regressive

NB: In practice, selection of a working correlation structure should be guided by a priori knowledge and/or exploratory analysis

geepack

- geepack implements estimating equations for β, α, and ϕ
- geeglm
  - Syntax similar to glm; returns an object similar to a glm object
  - An anova method provides multivariate Wald tests for joint hypotheses
  - Calls a fitter function geese to solve the estimating equations
- geese
  - Provides estimation and inference for β, α, and ϕ
  - Model objects are available within geeglm objects

```
names(m1)
names(m1$geese)
m1$geese$vbeta
```

R commands

```
load("ichs.RData")

library(geepack)

m1 <- geeglm(infection ~ xerop + age + gender + hfora + cost + sint,
             id=id, data=ichs, family="binomial", corstr="independence")

m2 <- geeglm(infection ~ xerop + age + gender + hfora + cost + sint,
             id=id, data=ichs, family="binomial", corstr="exchangeable")

m3 <- geeglm(infection ~ xerop + age + gender + hfora + cost + sint,
              id=id, data=ichs, family="binomial", corstr="ar1")
```

GEE-independence

```
> summary(m1)
 Coefficients:
            Estimate  Std.err   Wald Pr(>|W|)
(Intercept) -2.42134  0.16907 205.10  < 2e-16 ***
xerop        0.73148  0.42246   3.00  0.08337 .
age         -0.03188  0.00624  26.08  3.3e-07 ***
gender      -0.39364  0.23571   2.79  0.09492 .
hfora       -0.04944  0.02467   4.01  0.04511 *
cost        -0.58029  0.16928  11.75  0.00061 ***
sint        -0.16536  0.14865   1.24  0.26595
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Estimated Scale Parameters:
            Estimate Std.err
(Intercept)     1.02   0.644

Correlation: Structure = independence
Number of clusters:   275   Maximum cluster size: 6
```

GEE-exchangeable

```
> summary(m2)
 Coefficients:
            Estimate  Std.err   Wald Pr(>|W|)
(Intercept) -2.39852  0.17033 198.30  < 2e-16 ***
xerop        0.62693  0.43618   2.07  0.15063
age         -0.03162  0.00627  25.44  4.6e-07 ***
gender      -0.41887  0.23631   3.14  0.07631 .
hfora       -0.05282  0.02464   4.60  0.03205 *
cost        -0.57171  0.16846  11.52  0.00069 ***
sint        -0.16208  0.14556   1.24  0.26550
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Estimated Scale Parameters:
            Estimate Std.err
(Intercept)     1.02   0.655

Correlation: Structure = exchangeable  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha   0.0452  0.0449
Number of clusters:   275   Maximum cluster size: 6
```

GEE-AR(1)

```
> summary(m3)
 Coefficients:
            Estimate  Std.err   Wald Pr(>|W|)
(Intercept) -2.41535  0.16926 203.64  < 2e-16 ***
xerop        0.66981  0.44020   2.32  0.12810
age         -0.03197  0.00625  26.13  3.2e-07 ***
gender      -0.39516  0.23579   2.81  0.09376 .
hfora       -0.05095  0.02464   4.28  0.03863 *
cost        -0.57446  0.16839  11.64  0.00065 ***
sint        -0.17108  0.14754   1.34  0.24624
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Estimated Scale Parameters:
            Estimate Std.err
(Intercept)     1.02   0.644

Correlation: Structure = ar1  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha   0.0526  0.0544
Number of clusters:   275   Maximum cluster size: 6
```

Results

	$\hat{\beta}_1$ (SE)	exp($\hat{\beta}_1$) (95% CI)
Independence	0.73 (0.42)	2.08 (0.91, 4.76)
Exchangeable	0.63 (0.44)	1.87 (0.80, 4.40)
Auto-regressive	0.67 (0.44)	1.95 (0.83, 4.63)

- Vitamin A deficient children have an increased risk of respiratory infection, but confidence interval includes the null-hypothesized value
- geese provides estimation and inference for β, α, and ϕ
- Cannot reject the hypothesis that $\alpha = 0$
- **Note**: Model fit can be evaluated using QIC (Pan, 2001)

Working correlation structures

Exchangeable :

$$\begin{bmatrix} 1 & & & & & \\ 0.045 & 1 & & & & \\ 0.045 & 0.045 & 1 & & & \\ 0.045 & 0.045 & 0.045 & 1 & & \\ 0.045 & 0.045 & 0.045 & 0.045 & 1 & \\ 0.045 & 0.045 & 0.045 & 0.045 & 0.045 & 1 \end{bmatrix}$$

Auto-regressive :

$$\begin{bmatrix} 1 & & & & & \\ 0.053 & 1 & & & & \\ 0.003 & 0.053 & 1 & & & \\ 0.000 & 0.003 & 0.053 & 1 & & \\ 0.000 & 0.000 & 0.003 & 0.053 & 1 & \\ 0.000 & 0.000 & 0.000 & 0.003 & 0.053 & 1 \end{bmatrix}$$

Stata commands

```
* Declare the dataset to be "panel" data, grouped by id
* with time variable time
xtset id time

* Fit models with an exchangeable correlation structure
xtgee infection xerop age gender hfora cost sint,
          family(binomial) link(logit) corr(exch) robust

* Examine working correlation structure
estat wcorr
```

GEE-exchangeable

```
GEE population-averaged model                   Number of obs      =      1200
Group variable:                        id       Number of groups   =       275
Link:                               logit       Obs per group: min =         1
Family:                          binomial                      avg =       4.4
Correlation:                 exchangeable                      max =         6
                                                Wald chi2(6)       =     41.27
Scale parameter:                        1       Prob > chi2        =    0.0000

                                  (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Semi-robust
   infection |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       xerop |   .6269335   .4369803     1.43   0.151    -.2295322    1.483399
         age |  -.0316238    .006281    -5.03   0.000    -.0439343   -.0193133
      gender |  -.4188661   .2367394    -1.77   0.077    -.8828669    .0451347
       hfora |  -.0528237   .0246853    -2.14   0.032    -.1012059   -.0044414
        cost |  -.5717089   .1687711    -3.39   0.001    -.9024942   -.2409237
        sint |   -.162076   .1458239    -1.11   0.266    -.4478856    .1237335
       _cons |   -2.39852   .1706357   -14.06   0.000     -2.73296    -2.06408
------------------------------------------------------------------------------
```

Working correlation structure

```
. estat wcorr

Estimated within-id correlation matrix R:

      |         c1          c2          c3          c4          c5          c6
------+------------------------------------------------------------------------
   r1 |          1
   r2 |  .0451627           1
   r3 |  .0451627   .0451627           1
   r4 |  .0451627   .0451627   .0451627           1
   r5 |  .0451627   .0451627   .0451627   .0451627           1
   r6 |  .0451627   .0451627   .0451627   .0451627   .0451627           1
```

Mixed effects models

Do vitamin A deficient children have an increased risk of infection?

$$\begin{aligned} \mu_{ij}^{\star} &= \mathrm{E}[Y_{ij} \mid \gamma_{0i}] \\ &= \mathrm{P}[Y_{ij} = 1 \mid \gamma_{0i}] \end{aligned}$$

$$\begin{aligned} \operatorname{logit} \mu_{ij}^{\star} &= \log \frac{\mu_{ij}^{\star}}{1 - \mu_{ij}^{\star}} \\ &= (\beta_0^{\star} + \gamma_{0i}) + \beta_1^{\star} \, \text{Xerophthalmia}_{ij} + \cdots \end{aligned}$$

for $i = 1, \ldots, 275$ and $j = 1, \ldots, m_i$

- $\exp \beta_1^{\star}$ represents the ratio of the expected odds of respiratory infection for a typical individual with vitamin A deficiency to that for a typical individual with a sufficient amount of vitamin A of the same age, gender, ...
- $\exp \beta_1^{\star}$ is therefore a **conditional** parameter
- Respiratory infection is rare so odds ratio approximates relative risk

R commands

- Use the `glmer` command in the `lme4` library

```
library(lme4)
?glmer

m_ri <- glmer(infection ~ (1 | id) + factor(xerop)
              + age + factor(gender) + hfora
              + cost + sint,
              family=binomial, data=ichs, nAGQ=7)

methods(class="merMod")
expit <- function(x){exp(x)/(1+exp(x))}
expit(fixef(m_ri)[1])
expit(fixef(m_ri)[1]-1.96*sqrt(VarCorr(m_ri)$id[[1]]))
expit(fixef(m_ri)[1]+1.96*sqrt(VarCorr(m_ri)$id[[1]]))
```

Random intercepts model

```
> summary(m_ri)

Random effects:
 Groups Name        Variance Std.Dev.
 id     (Intercept) 0.647    0.8044
Number of obs: 1200, groups:  id, 275

Fixed effects:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.64240    0.21202 -12.463  < 2e-16 ***
factor(xerop)1   0.63182    0.47991   1.317 0.187996
age             -0.03369    0.00727  -4.634 3.59e-06 ***
factor(gender)1 -0.43570    0.25741  -1.693 0.090528 .
hfora           -0.05479    0.02254  -2.431 0.015056 *
cost            -0.59871    0.17392  -3.442 0.000577 ***
sint            -0.16450    0.17463  -0.942 0.346180
```

Interpreting random effects components

- For continuous outcomes interpreting random effects is 'easy' because their standard deviation is on the scale of the outcome
- For binary outcomes the standard deviation is on the log-odds scale
- Recall for a GLMM with random intercepts

$$\gamma_{0i} \sim N(0,\, G_{11}) \Leftrightarrow (\beta_0^\star + \gamma_{0i}) \sim N(\beta_0^\star,\, G_{11})$$

- In the ICHS analysis the intercept corresponds to the log odds of respiratory infection among females, age 36 months, . . . , with a sufficient amount of vitamin A
- We can use $\hat{\beta}_0^\star$ and $\hat{G}_{11}$ to form an interval to quantify variability in the probability of respiratory infection across these individuals

$$\text{expit}(\hat{\beta}_0^\star \pm 1.96 \times \hat{G}_{11}) = \frac{\exp(\hat{\beta}_0^\star \pm 1.96 \times \hat{G}_{11})}{1 + \exp(\hat{\beta}_0^\star \pm 1.96 \times \hat{G}_{11})},$$

  which is calculated to be 0.07 (0.01, 0.26)
- **NB**: This is **not** a confidence interval for $\beta_0^\star$

Conditional and marginal effects

- Parameter estimates obtained from a **marginal** model (as obtained via a GEE) estimate **population-averaged** contrasts
- Parameter estimates obtained from a **conditional** model (as obtained via a GLMM) estimate **subject-specific** contrasts
- In a linear model for a Gaussian outcome with an identity link these contrasts are equivalent; not the case with non-linear models
  - Depends on the outcome distribution
  - Depends on the specified random effects

Conditional and marginal effects

		Fitted conditional model	
Outcome	Coefficient	Random intercept	Random intercept/slope
Continuous	Intercept	Marginal	Marginal
	Slope	Marginal	Marginal
Count	Intercept	Conditional	Conditional
	Slope	Marginal	Conditional
Binary	Intercept	Conditional	Conditional
	Slope	Conditional	Conditional

⋆ Marginal = population-averaged; conditional = subject-specific

Stata commands

```
* Declare the dataset to be "panel" data, grouped by id
* with time variable time
xtset id time

* Fit a model with random intercepts
help melogit
melogit infection i.xerop age i.gender hfora cost sint || id:

* Obtain predicted probabilities of infection,
* setting the random effects to 0
margins i.xerop, predict(mu fixed)
```

Random intercepts model

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

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

Integration method: mvaghermite                 Integration points =         7

                                                Wald chi2(6)       =     35.62
Log likelihood = -334.75137                     Prob > chi2        =    0.0000
------------------------------------------------------------------------------
   infection |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
             |
     1.xerop |   .6317689   .4799255     1.32   0.188    -.3088678    1.572406
         age |  -.0336883   .0072704    -4.63   0.000    -.0479379   -.0194386
    1.gender |  -.4357064   .2574121    -1.69   0.091    -.9402248     .068812
       hfora |  -.0547912   .0225386    -2.43   0.015    -.0989661   -.0106164
        cost |   -.598695   .1739193    -3.44   0.001    -.9395706   -.2578193
        sint |  -.1644847   .1746269    -0.94   0.346     -.506747    .1777777
       _cons |  -2.642403   .2120549   -12.46   0.000    -3.058023   -2.226783
------------------------------------------------------------------------------
```

Random intercepts model

```
-------------+----------------------------------------------------------------
id           |
   var(_cons)|   .6470842   .3492486                      .2246697    1.863704
-------------------------------------------------------------------------------
LR test vs. logistic regression: chibar2(01) =     5.52 Prob>=chibar2 = 0.0094


Predictive margins                                Number of obs   =       1200
Model VCE    : OIM

Expression   : Predicted mean, fixed portion only, predict(mu fixed)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       xerop |
          0  |   .0709704   .0106353     6.67   0.000     .0501256    .0918152
          1  |   .1224475   .0496301     2.47   0.014     .0251743    .2197208
------------------------------------------------------------------------------
```

Summary

- Exploratory analysis with binary outcomes is not straightforward
  - Plots of raw data not always useful
  - Aggregated percents (means) can summarize mean response
  - Correlation can be examined using correlations or odds ratios
- GEE provides marginal, population-averaged contrasts
  - Ratio of the expected odds of respiratory infection among a population of vitamin A deficient children to that for a population of children replete with vitamin A of the same age, gender, ...
- GLMM provides conditional, subject-specific contrasts
  - Ratio of the expected odds of respiratory infection for a typical individual with vitamin A deficiency to that for a typical individual with a sufficient amount of vitamin A of the same age, gender, ...
  - Random effects variance components quantify heterogeneity in effects
- Lack of significance likely due to small number of exposed cases