Longitudinal Data Analysis

CATEGORICAL RESPONSE DATA

Motivation

- Vaccine preparedness study (VPS), 1995-1998.
  - 5,000 subjects with high-risk for HIV acquisition.
  - Feasibility of phase III HIV vaccine trials.
  - Willingness, knowledge?

Motivation

- VPS Informed Consent Substudy (IC)
  - 20% selected to undergo mock informed consent.
  - Understanding of key items at 6mo, 12mo, 18mo.
- **Reference**: Coletti et al. (2003) *JAIDS*

Simple Example: VPS IC Analysis

To develop methods which assure that participants in future HIV vaccine trials understand the implications and potential risks of participating, the HIVNET developed a prototype informed consent process for a hypothetical future HIV vaccine efficacy trial. A 20% random subsample of the 4,892 Vaccine Preparedness Study (VPS) cohort was enrolled in a mock informed consent process at month 3 of the study (between the enrollment visit and the scheduled follow-up visit at month 6). Knowledge of 10 key HIV concepts and willingness to participate in future vaccine efficacy trials among these participants were compared with knowledge and willingness levels of participants not randomized to the informed consent procedure.

Simple Example: VPS IC Analysis

Items:

- Q4SAFE – "We can be sure that the HIV vaccine is safe once we begin phase III testing"

- NURSE – "The study nurse decides whether placebo or active product is given to a participant"

EDA – time cross-sectional

Baseline

```
ICgroup |q4safe0
        |0      |1      |RowTotl|
 -------+-------+-------+-------+
 0      |218    |282    |500    |
        |0.44   |0.56   |       |
 -------+-------+-------+-------+
 1      |216    |284    |500    |
        |0.43   |0.57   |       |
 -------+-------+-------+-------+
```

EDA – time cross-sectional

Post-Intervention, +3 months

```
ICgroup |q4safe6
        |0      |1      |RowTotl|
 -------+-------+-------+-------+
 0      |226    |274    |500    |
        |0.45   |0.55   |       |
 -------+-------+-------+-------+
 1      |180    |320    |500    |
        |0.36   |0.64   |       |
 -------+-------+-------+-------+
```

EDA – time cross-sectional

Post-Intervention, +9 months

```
ICgroup |q4safe12
        |0       |1       |RowTotl|
 -------+--------+--------+-------+
 0      |208     |292     |500    |
        |0.42    |0.58    |       |
 -------+--------+--------+-------+
 1      |177     |323     |500    |
        |0.35    |0.65    |       |
 -------+--------+--------+-------+
```

Regression Models

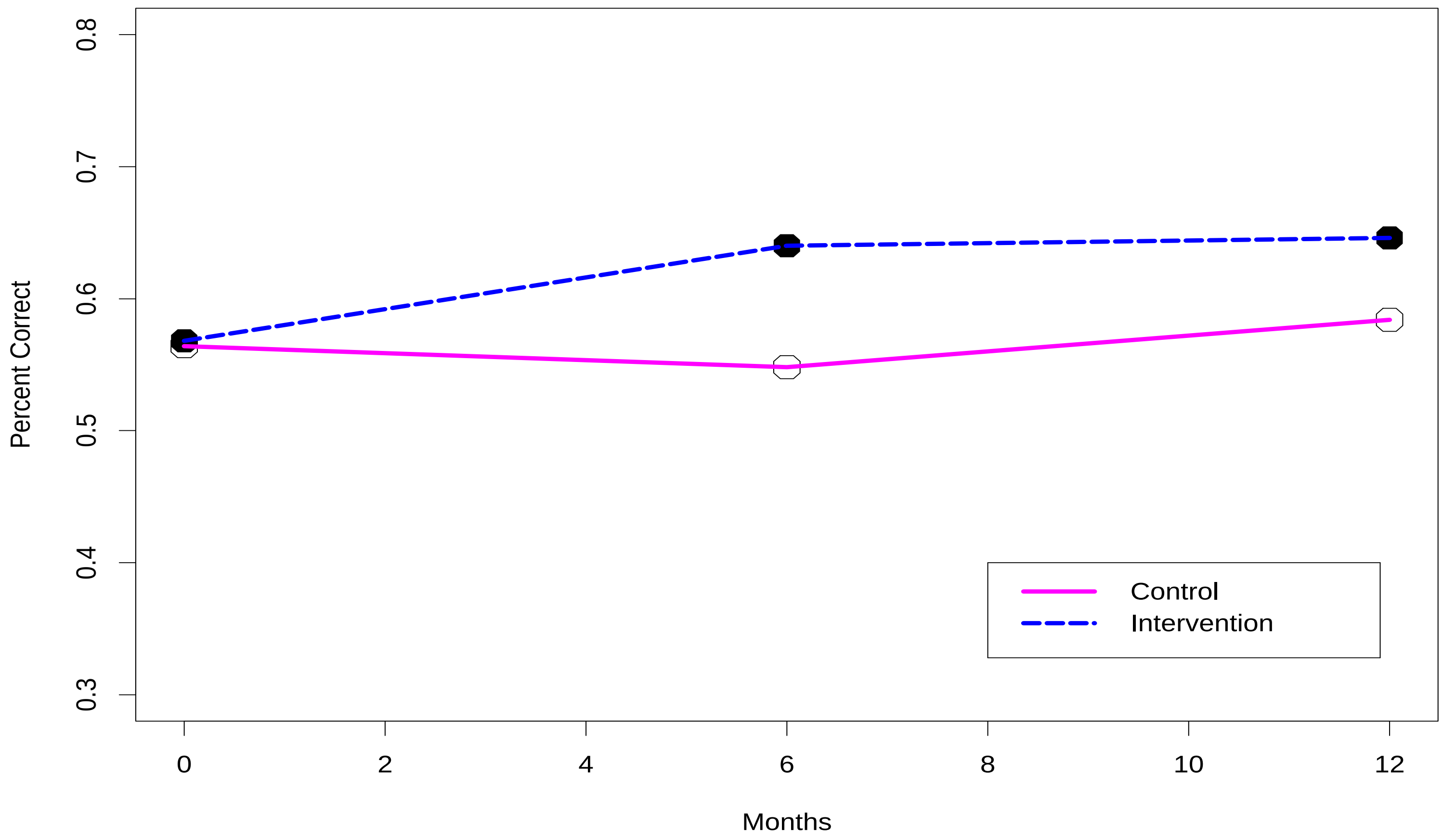
Q: Is there an intervention effect? If so what is it?

Q: Does the intervention effect "wane"?

Regression Models:

$$
\begin{aligned}
Y_{ij} &= \text{response at time } j \text{ for subject } i \\
\mu_{ij} &= E(Y_{ij} \mid X_{ij})
\end{aligned}
$$

HIVNET IC – Percent by Time and Group



Regression Models

Regression Models:

$$
\begin{aligned}
\text{logit}(\mu_{ij}) \quad = \quad & \beta_0 \;+\; \beta_1 \cdot (\text{Tx}) \;+ \\
& \beta_2 \cdot (\text{Time}{=}6) \;+\; \beta_3 \cdot (\text{Time}{=}12) \;+ \\
& \beta_4 \cdot (\text{Time}{=}6 \cdot \text{Tx}) \;+\; \beta_5 \cdot (\text{Time}{=}12 \cdot \text{Tx})
\end{aligned}
$$

Regression Models

Analysis Options:

- Cross-sectional analyses at 0, 6, and 12 month.

⋆ **Semi-parametric methods (GEE)**

- "Random effects" models. / Transition models.

Longitudinal Data Analysis

GENERALIZED ESTIMATING EQUATIONS (GEE)

GEE Liang and Zeger (1986)

Q: We've seen that the LMM assuming multivariate normality can be used for likelihood based estimation with continuous response variables. What about models/methods for discrete response variables such as binary data?

A: There are semi-parametric approaches (GEE) and likelihood based methods (GLMMs and other models).

GEE Liang and Zeger (1986)

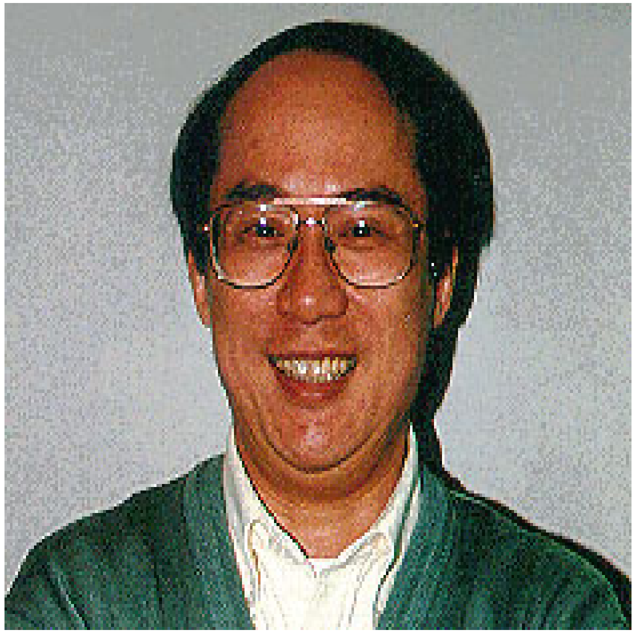
$\boxed{\star\star\star}$ Let's consider GEE first:

- Focus on a generalized linear model regression parameter that characterizes systematic variation across covariate levels: $\boldsymbol{\beta}$.

- Repeated measurements, clustered data, multivariate response.

- Correlation structure is a $nuisance$ feature of the data.

Liang and Zeger (not 1986)



Professor JHU
Vice President NHRI, Taiwan

Chair Biostatistics JHU

GEE1 - Notation

Data:

$$
\begin{array}{rcl}
Y_{i1}, Y_{i2}, \ldots, Y_{ij}, \ldots, Y_{in_i} & & \text{response variables} \\
\boldsymbol{X}_{i1}, \boldsymbol{X}_{i2}, \ldots, \boldsymbol{X}_{ij}, \ldots, \boldsymbol{X}_{in_i} & & \text{covariate vectors} \\
 & & \\
i \in [1, N] & : & \text{index for cluster / subject} \\
j \in [1, n_i] & : & \text{index for measurement} \\
 & & \quad \text{within cluster}
\end{array}
$$

GEE1 - Notation

Assumptions:

- Measurements are independent across clusters (can be relaxed for time and space).

- Measurements may be correlated within cluster.

Mean Model: (primary focus of analysis)

$$E[Y_{ij} \mid \boldsymbol{X}_{ij}] = \mu_{ij}$$

$$\begin{aligned} g(\mu_{ij}) &= \beta_0 + \beta_1 \cdot X_{ij,1} + \ldots + \beta_p \cdot X_{ij,p} \\ &= \boldsymbol{X}_{ij}\boldsymbol{\beta} \end{aligned}$$

Marginal Mean

Mean Model: (primary focus of analysis)

$$
\begin{aligned}
E[Y_{ij} \mid \boldsymbol{X}_{ij}] &= \mu_{ij} \\
g(\mu_{ij}) &= \boldsymbol{X}_{ij}\boldsymbol{\beta}
\end{aligned}
$$

This can be any generalized linear model. For example,

$$
\begin{aligned}
P[Y_{ij} = 1 \mid \boldsymbol{X}_{ij}] &= \pi_{ij} \\
\log(\frac{\pi_{ij}}{1 - \pi_{ij}}) &= \boldsymbol{X}_{ij}\boldsymbol{\beta}
\end{aligned}
$$

Q: Why is this a **marginal** mean?

Marginal Mean

A: There's no extra variable(s) that we condition on (like in some other models for multivariate data).

- Log-linear models: $E[\, Y_{ij} \mid Y_{ik},\ k \neq j,\ \boldsymbol{X}_{ij}]$

- Transition models: $E[\, Y_{ij} \mid Y_{ik},\ k < j,\ \boldsymbol{X}_{ij}]$

- Latent variable models: $E[Y_{ij} \mid b_{ij},\ \boldsymbol{X}_{ij}]$

GEE - covariance

Q: But what about the fact that data are clustered?

A: Choose a Correlation Model: (nuisance)

$$\begin{aligned} \text{var}(Y_{ij} \mid \boldsymbol{X}_i) &= V_{ij} \\ \boldsymbol{A}_i &= \text{diag}(V_{ij}) \end{aligned}$$

$$\begin{aligned} \text{corr}(Y_{ij}, Y_{ik} \mid \boldsymbol{X}_i) &= \rho_{ijk}(\boldsymbol{\alpha}) \\ \boldsymbol{R}_i(\boldsymbol{\alpha}) &= \text{correlation matrix} \\ \boldsymbol{V}_i(\boldsymbol{\alpha}) &= \text{cov}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) \\ &= \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{A}_i^{1/2} \end{aligned}$$

- In GLMs V_{ij} is a function of the mean μ_{ij} [e.g. $\mu_{ij}(1 - \mu_{ij})$].
- The parameter $\boldsymbol{\alpha}$ characterizes the correlation.

GEE1 - Common Correlation Models

Independence:

$$\boldsymbol{R}_i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Exchangeable / equicorrelation:

$$\boldsymbol{R}_i(\alpha) = \begin{bmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha \\ \alpha & \alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & 1 \end{bmatrix}$$

Unstructured:

$$\boldsymbol{R}_i(\boldsymbol{\alpha}) = \begin{bmatrix} 1 & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & 1 & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & 1 & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & 1 \end{bmatrix}$$

GEE1 - Common Correlation Models

AR-1:

$$\boldsymbol{R}_i(\alpha) = \begin{bmatrix} 1 & \alpha^1 & \alpha^2 & \alpha^3 \\ \alpha^1 & 1 & \alpha^1 & \alpha^2 \\ \alpha^2 & \alpha^1 & 1 & \alpha^1 \\ \alpha^3 & \alpha^2 & \alpha^1 & 1 \end{bmatrix}$$

Stationary m-dependent $(m = 2)$:

$$\boldsymbol{R}_i(\boldsymbol{\alpha}) = \begin{bmatrix} 1 & \alpha_1 & \alpha_2 & 0 \\ \alpha_1 & 1 & \alpha_1 & \alpha_2 \\ \alpha_2 & \alpha_1 & 1 & \alpha_1 \\ 0 & \alpha_2 & \alpha_1 & 1 \end{bmatrix}$$

Non-stationary m-dependent ($m = 2$):

$$\boldsymbol{R}_i(\boldsymbol{\alpha}) = \begin{bmatrix} 1 & \alpha_{12} & \alpha_{13} & 0 \\ \alpha_{21} & 1 & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & 1 & \alpha_{34} \\ 0 & \alpha_{42} & \alpha_{43} & 1 \end{bmatrix}$$

GEE1 - semiparametric model

Q: Does specification of a mean model, $\mu_{ij}(\boldsymbol{\beta})$, and a correlation model, $\boldsymbol{R}_i(\boldsymbol{\alpha})$, identify a complete probability model for $\boldsymbol{Y}_i$?

- No.

- If further assumptions can be made then a probability model can be identified. In general, for categorical data this is a difficult task.

- The model $\{\mu_{ij}(\boldsymbol{\beta}), \boldsymbol{R}_i(\boldsymbol{\alpha})\}$ is *semiparametric* since it only specifies the first two multivariate moments (mean and covariance) of $\boldsymbol{Y}_i$.

GEE1 - semiparametric model

Q: Without a likelihood function how can we estimate $\boldsymbol{\beta}$ (and possibly $\boldsymbol{\alpha}$) and perform valid statistical inference that takes the dependence into consideration?

A: Construct an unbiased estimating function.

GEE1 - estimation

Define:

$$
\begin{aligned}
\boldsymbol{D}_i(\boldsymbol{\beta}) &= \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \\
\boldsymbol{D}_i(j,k) &= \frac{\partial \mu_{ij}}{\partial \beta_k} \\
\boldsymbol{V}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{A}_i^{1/2}
\end{aligned}
$$

Define:

$$
U(\boldsymbol{\beta}) = \sum_{i=1}^{N} \boldsymbol{D}_i^T(\boldsymbol{\beta}) \boldsymbol{V}_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \left\{ \boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}) \right\}
$$

Note:

- $U(\boldsymbol{\beta})$ is called an estimating function.
- $U(\boldsymbol{\beta})$ also depends on the model/value for $\boldsymbol{\alpha}$.

Estimating Equations: solution to the following system of equations defines an estimator $\widehat{\boldsymbol{\beta}}$

$$\begin{aligned} \mathbf{0} &= U(\widehat{\boldsymbol{\beta}}) \\ &= \sum_{i=1}^{N} \boldsymbol{D}_i^T(\boldsymbol{\beta}) \boldsymbol{V}_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \left\{ \boldsymbol{Y}_i - \boldsymbol{\mu}_i(\widehat{\boldsymbol{\beta}}) \right\} \end{aligned}$$

Note: use $\boldsymbol{D}_i$, and $\boldsymbol{V}_i(\boldsymbol{\alpha})$ to denote $\boldsymbol{D}_i(\boldsymbol{\beta})$ and $\boldsymbol{V}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$.

Estimating Equations

$$\mathbf{0} = \sum_{i=1}^{N} \underbrace{\boldsymbol{D}_i^T(\boldsymbol{\beta})}_{\boxed{3}} \underbrace{\boldsymbol{V}_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha})}_{\boxed{2}} \underbrace{[\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})]}_{\boxed{1}}$$

- $\boxed{1}$ – The model for the mean, $\mu_i(\boldsymbol{\beta})$, is compared to the observed data, $\boldsymbol{Y}_i$. Setting the equations to equal $\mathbf{0}$ tries to minimize the difference between **observed** and **expected**.

- $\boxed{2}$ – Estimation uses the inverse of the variance (covariance) to weight the data from subject i. Thus, more weight is given to differences between observed and expected for those subjects who contribute more information.

- $\boxed{3}$ – This is simply a "change of scale" from the scale of the mean, μ_i, to the scale of the regression coefficients (covariates).

GEE1 - estimation

Q: What are the properties of $\widehat{\boldsymbol{\beta}}$, the regression estimate?

Robustness Property:

- The regression coefficient estimate, $\widehat{\boldsymbol{\beta}}$, will be correct (in large samples) **even if** you choose the wrong dependence model.

- However, the variance of the regression estimate must capture the correlation in the data, either through choosing the correct correlation model, or via an alternative variance estimate.

- Choosing a "wise" (approximately correct) correlation model will make the regression estimate $\widehat{\boldsymbol{\beta}}$ more efficient in the extraction of information (ie. $\widehat{\boldsymbol{\beta}}$ has smallest variance if correct correlation model).

GEE and Standard Error Estimates

GEE Specification

(1) A flexible regression model for the mean response (linear, logistic).

(2) A correlation model (independence, exchangeable).

Q: What if the selected correlation model is not correct?

GEE and Standard Error Estimates

A: GEE also computes a **sandwich variance** estimator.

$\Rightarrow$ a.k.a. "empirical variance"

$\Rightarrow$ a.k.a. "robust variance"

$\Rightarrow$ a.k.a. "Huber-White correction"

$\boxed{\star}$ The empirical variance gives valid standard errors for the estimated regression coefficients <u>even if</u> the correlation model was wrong.

- The empirical variance is valid in "large samples" – this means it can be used with data sets that contain at least 40 subjects.

Empirical Standard Errors

- On page 160 we considered weighted least squares regression estimates and stated that when a weight, $\boldsymbol{W}_i$ is used that is <u>not</u> equal to the inverse of the variance (covariance) then:

$$\boldsymbol{W}_i \neq \boldsymbol{\Sigma}_i^{-1} \quad \Rightarrow$$

$$\text{var}\left[\widehat{\boldsymbol{\beta}}(\boldsymbol{W})\right] = \overbrace{\boldsymbol{A}^{-1}}^{\textbf{bread}} \underbrace{\left(\sum_i \boldsymbol{X}_i^T \boldsymbol{W}_i \,\text{var}(\boldsymbol{Y}_i)\, \boldsymbol{W}_i \boldsymbol{X}_i\right)}_{\textbf{cheese}} \overbrace{\boldsymbol{A}^{-1}}^{\textbf{bread}}$$

$$\boldsymbol{A} = \sum_i \boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{X}_i$$

- **Q**: What to do about not having a correct model for var$(\boldsymbol{Y}_i)$?

Empirical Standard Errors

- **A**: We can try to estimate the middle part of this sandwich variance estimate, and then would have a valid estimate of the standard error.
- Try the simplest idea:

$$\widehat{\text{var}}\left[\widehat{\beta}(\boldsymbol{W})\right] = \overbrace{\boldsymbol{A}^{-1}}^{\textbf{bread}} \underbrace{\left(\sum_i \boldsymbol{X}_i^T \boldsymbol{W}_i \, (\boldsymbol{Y}_i - \mu_i)^2 \, \boldsymbol{W}_i \boldsymbol{X}_i\right)}_{\textbf{cheese}} \overbrace{\boldsymbol{A}^{-1}}^{\textbf{bread}}$$

- Where we use $(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)^2$, or the vector version of the variance (covariance) $(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)^T$ to estimate the variance (covariance).

Empirical Standard Errors

- This idea works since we actually use the sum (average) of these estimates where we sum (average) over the subjects in the data.
  - ▷ No single variance is estimated very well.
  - ▷ But the **average** or total variance is estimated well!
- For generalized linear models (logistic, poisson) this same basic idea is used.
- $\boxed{\text{Implication}}$ when using empirical s.e.
  - ▷ $\widehat{\beta}_k /$ s.e. – valid test
  - ▷ $\widehat{\beta}_k \pm 1.96 \times$ s.e. – valid confidence interval
- Inference using the **empirical** (robust) standard errors is correct inference even when a poor choice is made for the correlation model.

GEE – Summary

Models

- **Mean model** = general regression model. Focus of analysis.
- **Correlation model** = simple choices. Nuisance.

GEE – Summary

Estimates

- **Regression estimate**, $\widehat{\boldsymbol{\beta}}$.
  - Valid estimate regardless of correlation choice.
  - Correlation choice wrong $\Rightarrow$ $\widehat{\boldsymbol{\beta}}$ still o.k.
- **Standard error estimates**.
  - Model-based standard errors.
    - $\star$ If correlation choice is correct $\Rightarrow$ valid.
  - Empirical standard errors.
    - $\star$ If correlation choice is <u>incorrect</u> $\Rightarrow$ still valid!

Example: Informed Consent Analysis

- Compare intervention groups, IC=yes to IC=no, separately at month 0, month 6, and month 12.

  $\Rightarrow$ Repeat cross-sectional analyses.

- Use GEE to analyze all follow-up times.

- Consider the question of treatment "waning".

  $\Rightarrow$ compare effects at 6mo and 12mo.

STATA Analysis Program

```
**************************************************************
* HivnetIC.do                                                *
**************************************************************
*                                                            *
* PURPOSE:  analysis of HIVNET Informed Consent Data         *
*                                                            *
* AUTHOR:  P. Heagerty                                       *
*                                                            *
* DATE:  02 May 2005                                         *
**************************************************************

infile id group education age cohort ICgroup will0 know0 ///
       q4safe0 q4safe6 q4safe12 ///
       nurse0 nurse6 nurse12 using HivnetWide.dat

***
*** recode and label variables
***

gen knowhigh = know0
recode knowhigh min/7=0 8/max=1
```

```
(EDITED)

***
*** univariate summaries
***
tabulate q4safe0
tabulate q4safe6
tabulate q4safe12

***
*** bivariate summaries
***
tabulate ICgroup q4safe0, row chi
logit q4safe0 ICgroup

tabulate ICgroup q4safe6, row chi
logit q4safe6 ICgroup

tabulate ICgroup q4safe12, row chi
logit q4safe12 ICgroup

***
*** correlation
***
```

```
sort ICgroup
by ICgroup: corr q4safe0 q4safe6 q4safe12

***
*** transitions
***
tabulate q4safe0 q4safe6, row chi

tabulate q4safe6 q4safe12, row chi
```

Cross-sectional Results **Baseline**

```
. tabulate ICgroup q4safe0, row chi

           |       q4safe0
   ICgroup |         0          1 |     Total
-----------+----------------------+----------
         0 |       218        282 |       500
           |     43.60      56.40 |    100.00
-----------+----------------------+----------
         1 |       216        284 |       500
           |     43.20      56.80 |    100.00
-----------+----------------------+----------
     Total |       434        566 |     1,000
           |     43.40      56.60 |    100.00

          Pearson chi2(1) =   0.0163   Pr = 0.898
```

Cross-sectional Results **Baseline**

```
. logit q4safe0 ICgroup

Logit estimates

Log likelihood = -684.40156
------------------------------------------------------------------------
 q4safe0 |    Coef.   Std. Err.      z     P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------
 ICgroup |  0.01628   .127608      0.13   0.898    -.23382     .26639
   _cons |  0.25741   .090184      2.85   0.004     .08065     .43417
------------------------------------------------------------------------
```

Cross-sectional Results **Month 6**

```
. tabulate ICgroup q4safe6, row chi

           |        q4safe6
   ICgroup |         0          1 |     Total
-----------+----------------------+----------
         0 |       226        274 |       500
           |     45.20      54.80 |    100.00
-----------+----------------------+----------
         1 |       180        320 |       500
           |     36.00      64.00 |    100.00
-----------+----------------------+----------
     Total |       406        594 |     1,000
           |     40.60      59.40 |    100.00

          Pearson chi2(1) =   8.7741   Pr = 0.003
```

Cross-sectional Results **Month 6**

```
. logit q4safe6 ICgroup

Logit estimates

Log likelihood = -670.97514
------------------------------------------------------------------------
 q4safe6 |    Coef.    Std. Err.       z     P>|z|      [95% Conf. Interval]
---------+--------------------------------------------------------------
 ICgroup |    0.38277     .129441      2.96    0.003      .12907     .63647
   _cons |    0.19259     .089857      2.14    0.032      .01647     .36871
------------------------------------------------------------------------
```

Cross-sectional Results **Month 12**

```
. tabulate ICgroup q4safe12, row chi

           |       q4safe12
   ICgroup |         0          1 |     Total
-----------+----------------------+----------
         0 |       208        292 |       500
           |     41.60      58.40 |    100.00
-----------+----------------------+----------
         1 |       177        323 |       500
           |     35.40      64.60 |    100.00
-----------+----------------------+----------
     Total |       385        615 |     1,000
           |     38.50      61.50 |    100.00

          Pearson chi2(1) =   4.0587   Pr = 0.044
```

Cross-sectional Results **Month 12**

```
. logit q4safe12 ICgroup

Logit estimates

Log likelihood = -664.42786
------------------------------------------------------------------------
q4safe12 |    Coef.    Std. Err.      z     P>|z|    [95% Conf. Interval]
---------+--------------------------------------------------------------
 ICgroup |    0.26228    .13029      2.01   0.044     .00690      .51766
   _cons |    0.33921    .09073      3.74   0.000     .16138      .51704
------------------------------------------------------------------------
```

Correlations

```
-> ICgroup = 0
(obs=500)
             |  q4safe0  q4safe6 q4safe12
-------------+---------------------------
     q4safe0 |   1.0000
     q4safe6 |   0.4008   1.0000
    q4safe12 |   0.2480   0.3423   1.0000

---------------------------------------------------------------------
-> ICgroup = 1
(obs=500)
             |  q4safe0  q4safe6 q4safe12
-------------+---------------------------
     q4safe0 |   1.0000
     q4safe6 |   0.3385   1.0000
    q4safe12 |   0.3000   0.4381   1.0000
```

STATA Analysis Program

```
*************************************************************
*** create "long" format data                             ***
*************************************************************

*** this command takes variables that end in numbers (times),
*** such as q4safe0 q4safe6 q4safe12 and then "stacks" these
*** into a single variable (truncating the numbers from the names)
*** and creating a new variable which records the truncated numbers,
*** or times for the outcome.

reshape long q4safe, i(id) j(month)

list id q4safe month ICgroup education in 1/8
```

Reshaping the data

```
. reshape long q4safe, i(id) j(month)
(note: j = 0 6 12)

Data                                  wide   ->   long
-------------------------------------------------------------
Number of obs.                        1000   ->    3000
Number of variables                     19   ->      18
j variable (3 values)                        ->   month
xij variables:
               q4safe0 q4safe6 q4safe12   ->   q4safe
-------------------------------------------------------------
. list id q4safe month ICgroup education in 1/8
     +-------------------------------------------+
     | id   q4safe   month   ICgroup   educat~n |
     |-------------------------------------------|
  1. | 10        0       0         0          3 |
  2. | 10        0       6         0          3 |
  3. | 10        0      12         0          3 |
     |-------------------------------------------|
  4. | 13        0       0         1          3 |
  5. | 13        0       6         1          3 |
  6. | 13        0      12         1          3 |
     |-------------------------------------------|
  7. | 23        1       0         0          5 |
  8. | 23        0       6         0          5 |
     +-------------------------------------------+
```

STATA Analysis Program

```
*****************************************************************
*** GEE Analysis                                              ***
*****************************************************************

gen month6 = (month==6)
gen ICgroupXmonth6 = month6 * ICgroup

gen month12 = (month==12)
gen ICgroupXmonth12 = month12 * ICgroup


*** [1] Baseline and Month 6 Only

xtgee q4safe ICgroup month6 ICgroupXmonth6 if month<=6, ///
  i(id) corr(exchangeable) family(binomial) link(logit)

xtgee q4safe ICgroup month6 ICgroupXmonth6 if month<=6, ///
  i(id) corr(exchangeable) family(binomial) link(logit) robust

xtcorr
```

GEE Results for month 0 and month 6 **exchangeable**

```
. xtgee q4safe ICgroup month6 ICgroupXmonth6 if month<=6, ///
   i(id) corr(exchangeable) family(binomial) link(logit)

GEE population-averaged model
Group variable:                                    id
Link:                                           logit
Family:                                      binomial
Correlation:                             exchangeable
------------------------------------------------------------------------
      q4safe |   Coef.    Std. Err.     z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------
     ICgroup |  0.01628    .12760     0.13   0.898    -.23382     .26639
      month6 | -0.06481    .10107    -0.64   0.521    -.26292     .13328
ICgroupXmo~6 |  0.36648    .14432     2.54   0.011     .08362     .64935
       _cons |  0.25741    .09018     2.85   0.004     .08065     .43417
------------------------------------------------------------------------
```

GEE Results for month 0 and month 6 exchangeable / robust

```
. xtgee q4safe ICgroup month6 ICgroupXmonth6 if month<=6, ///
   i(id) corr(exchangeable) family(binomial) link(logit) robust

GEE population-averaged model
Link:                                       logit
Family:                                  binomial
Correlation:                         exchangeable
                         (standard errors adjusted for clustering on id)
------------------------------------------------------------------------
             |           Semi-robust
      q4safe |    Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------
     ICgroup |  0.01628   .12767      0.13   0.899    -.23395    .26651
      month6 | -0.06481   .09859     -0.66   0.511    -.25805    .12842
ICgroupXmo~6 |  0.36648   .14446      2.54   0.011     .08334    .64962
       _cons |  0.25741   .09022      2.85   0.004     .08056    .43425
------------------------------------------------------------------------
```

```
. xtcorr

Estimated within-id correlation matrix R:

         c1       c2
r1  1.0000
r2  0.3697  1.0000
```

STATA Analysis Program

```
*** [2] Baseline, Month 6, and Month 12

xtgee q4safe ICgroup month6 month12 ICgroupXmonth6 ICgroupXmonth12, ///
  i(id) corr(unstructured) t(month) family(binomial) link(logit)

xtgee q4safe ICgroup month6 month12 ICgroupXmonth6 ICgroupXmonth12, ///
  i(id) corr(unstructured) t(month) family(binomial) link(logit) robust

xtcorr

test ICgroupXmonth6 ICgroupXmonth12

test ICgroup ICgroupXmonth6 ICgroupXmonth12

lincom ICgroupXmonth12 - ICgroupXmonth6
```

HIVNET IC Regression

group	month0	month6	month12
control	β_0	$\beta_0 + \beta_{\text{month6}}$	$\beta_0 + \beta_{\text{month12}}$
intervention	β_0 $+\beta_{\text{ICgroup}}$	$\beta_0 + \beta_{\text{month6}}$ $+\beta_{\text{ICgroup}}$ $+\beta_{\text{ICgroup:month6}}$	$\beta_0 + \beta_{\text{month12}}$ $+\beta_{\text{ICgroup}}$ $+\beta_{\text{ICgroup:month12}}$

HIVNET IC Regression

- $\boxed{\text{Change in log odds: Baseline to Month 6}}$
  - ▷ **Control**:
  - ▷ **Intervention**:
- $\boxed{\text{Change in log odds: Baseline to Month 12}}$
  - ▷ **Control**:
  - ▷ **Intervention**:

GEE Results for months 0, 6, 12 **Unstructured** / **robust**

```
. xtgee q4safe ICgroup month6 month12 ICgroupXmonth6 ICgroupXmonth12, ///
   i(id) corr(unstructured) t(month) family(binomial) link(logit) robust

GEE population-averaged model
Link:                                     logit
Family:                                binomial
Correlation:                       unstructured
                     (standard errors adjusted for clustering on id)
------------------------------------------------------------------------
             |          Semi-robust
      q4safe |   Coef.    Std. Err.      z     P>|z|    [95% Conf. Interval]
-------------+----------------------------------------------------------
     ICgroup |   0.01628   .12767     0.13   0.899    -.23395     .26651
      month6 |  -0.06481   .09859    -0.66   0.511    -.25805     .12842
     month12 |   0.08180   .11099     0.74   0.461    -.13573     .29934
ICgroupXmo~6 |   0.36648   .14446     2.54   0.011     .08334     .64962
ICgroupXm~12 |   0.24600   .15543     1.58   0.114    -.05864     .55065
       _cons |   0.25741   .09022     2.85   0.004     .08056     .43425
------------------------------------------------------------------------
```

```
. xtcorr

Estimated within-id correlation matrix R:

         c1      c2      c3
r1  1.0000
r2  0.3697  1.0000
r3  0.2740  0.3902  1.0000
```

GEE Results for months 0, 6, 12 **Unstructured**

```
. test ICgroupXmonth6 ICgroupXmonth12

 ( 1)  ICgroupXmonth6 = 0
 ( 2)  ICgroupXmonth12 = 0

           chi2(  2) =    6.49
         Prob > chi2 =    0.0389
.
. test ICgroup ICgroupXmonth6 ICgroupXmonth12

 ( 1)  ICgroup = 0
 ( 2)  ICgroupXmonth6 = 0
 ( 3)  ICgroupXmonth12 = 0

           chi2(  3) =   11.02
         Prob > chi2 =    0.0116
```

```
.
. lincom ICgroupXmonth12 - ICgroupXmonth6

 ( 1) - ICgroupXmonth6 + ICgroupXmonth12 = 0
------------------------------------------------------------------------------
  q4safe |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
     (1) |  -.1204842   .1433102    -0.84   0.401     -.401367    .1603987
------------------------------------------------------------------------------
```

STATA Analysis Program

```
***alternative parameterization

gen post = (month>0)
gen ICgroupXpost = post * ICgroup

xtgee q4safe ICgroup post month12 ICgroupXpost ICgroupXmonth12, ///
  i(id) corr(unstructured) t(month) family(binomial) link(logit) robust


*** ANCOVA type analysis

xtgee q4safe post month12 ICgroupXpost ICgroupXmonth12, ///
  i(id) corr(unstructured) t(month) family(binomial) link(logit) robust

test ICgroupXpost ICgroupXmonth12


***adjustment for baseline covariates

xi: xtgee q4safe ICgroup post month12 ICgroupXpost ICgroupXmonth12 ///
  msm cohort school i.agecat, ///
```

```
  i(id) corr(unstructured) t(month) family(binomial) link(logit) robust

xtcorr

test ICgroupXpost ICgroupXmonth12

test ICgroup ICgroupXpost ICgroupXmonth12
```

group	month0	month6	month12
control	β_0	$\beta_0 + \beta_{\textsf{post}}$	$\beta_0 + \beta_{\textsf{post}} + \beta_{\textsf{month12}}$
intervention	β_0 $+\beta_{\textsf{ICgroup}}$	$\beta_0 + \beta_{\textsf{post}}$ $+\beta_{\textsf{ICgroup}}$ $+\beta_{\textsf{ICgroup:post}}$	$\beta_0 + \beta_{\textsf{post}} + \beta_{\textsf{month12}}$ $+\beta_{\textsf{ICgroup}}$ $+\beta_{\textsf{ICgroup:post}}$ $+\beta_{\textsf{ICgroup:month12}}$

HIVNET IC Regression

- $\boxed{\text{Change in log odds: Baseline to Month 6}}$
  - ▷ **Control**:
  - ▷ **Intervention**:
- $\boxed{\text{Change in log odds: Month 6 to Month 12}}$
  - ▷ **Control**:
  - ▷ **Intervention**:

GEE Results for months 0, 6, 12 **Unstructured** / **robust**

```
. xtgee q4safe ICgroup post month12 ICgroupXpost ICgroupXmonth12, ///
   i(id) corr(unstructured) t(month) family(binomial) link(logit) robust

GEE population-averaged model
Correlation:                       unstructured
                      (standard errors adjusted for clustering on id)
------------------------------------------------------------------------
             |          Semi-robust
      q4safe |   Coef.    Std. Err.      z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------
     ICgroup |   0.01628   .12767      0.13   0.899    -.23395     .26651
        post |  -0.06481   .09859     -0.66   0.511    -.25805     .12842
     month12 |   0.14662   .10361      1.42   0.157    -.05645     .34970
ICgroupXpost |   0.36648   .14446      2.54   0.011     .08334     .64962
ICgroupXm~12 |  -0.12048   .14331     -0.84   0.401    -.40136     .16039
       _cons |   0.25741   .09022      2.85   0.004     .080561    .43425
------------------------------------------------------------------------
```

GEE Results for months 0, 6, 12 **Unstructured** / **robust**

```
. xi: xtgee q4safe ICgroup post month12 ICgroupXpost ICgroupXmonth12 ///
   msm cohort school i.agecat, ///
   i(id) corr(unstructured) t(month) family(binomial) link(logit) robust

GEE population-averaged model
Correlation:                      unstructured
                         (standard errors adjusted for clustering on id)
------------------------------------------------------------------------
             |          Semi-robust
      q4safe |   Coef.    Std. Err.      z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------
     ICgroup |   0.07638   .13494     0.57   0.571    -.18811     .34087
        post |  -0.07214   .10937    -0.66   0.509    -.28652     .14222
     month12 |   0.16315   .11501     1.42   0.156    -.06226     .38857
ICgroupXpost |   0.40736   .16065     2.54   0.011     .09248     .72224
ICgroupXm~12 |  -0.13368   .15935    -0.84   0.402    -.44602     .17864
         msm |   0.65603   .14271     4.60   0.000     .37631     .93576
      cohort |  -0.15267   .10343    -1.48   0.140    -.35540     .05004
      school |   0.88680   .13379     6.63   0.000     .62457    1.14904
```

```
  _Iagecat_1 |   0.10980   .11960     0.92   0.359    -.12460    .34422
  _Iagecat_2 |   0.23471   .13290     1.77   0.077    -.02577    .49521
       _cons |  -0.83223   .17682    -4.71   0.000   -1.17880   -.48565
-------------------------------------------------------------------------

. xtcorr
Estimated within-id correlation matrix R:
        c1      c2      c3
r1  1.0000
r2  0.3031  1.0000
r3  0.1946  0.3167  1.0000
```

GEE Results for months 0, 6, 12 **Unstructured** / **robust**

```
. test ICgroupXpost ICgroupXmonth12

 ( 1)  ICgroupXpost = 0
 ( 2)  ICgroupXmonth12 = 0

           chi2(  2) =    6.49
         Prob > chi2 =    0.0390

.
. test ICgroup ICgroupXpost ICgroupXmonth12

 ( 1)  ICgroup = 0
 ( 2)  ICgroupXpost = 0
 ( 3)  ICgroupXmonth12 = 0

           chi2(  3) =   15.09
         Prob > chi2 =    0.0017
```

SAS: GEE using GENMOD

```
options linesize=80 pagesize=60;

data hivnet;
  infile 'HivnetIC-SAS.data';
  input y month ICgroup id month6 month12 post riskgp
        educ age cohort;
run;

proc genmod data=hivnet descending;
       class id riskgp;
       model y = post ICgroup ICgroup*post /
                 dist=binomial link=logit;
       repeated subject=id / corrw type=ar;
run;

proc genmod data=hivnet descending;
       class id riskgp;
       model y = post ICgroup ICgroup*post /
                 dist=binomial link=logit;
       repeated subject=id / corrw type=un;
run;
```

GEE Results for months 0, 6, 12 “Generic Prelude”

```
                         The GENMOD Procedure

                          Model Information

               Data Set                     WORK.HIVNET
               Distribution                    Binomial
               Link Function                      Logit
               Dependent Variable                     y
               Observations Used                   3000


                           Response Profile

                    Ordered                 Total
                      Value      y      Frequency

                          1      1           1775
                          2      0           1225

PROC GENMOD is modeling the probability that y='1'.
```

Parameter Information

Parameter	Effect
Prm1	Intercept
Prm2	post
Prm3	month12
Prm4	ICgroup
Prm5	post*ICgroup
Prm6	month12*ICgroup

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	2994	4039.6091	1.3492
Scaled Deviance	2994	4039.6091	1.3492
Pearson Chi-Square	2994	3000.0000	1.0020
Scaled Pearson X2	2994	3000.0000	1.0020
Log Likelihood		-2019.8046	

The GENMOD Procedure

Algorithm converged.

Analysis Of Initial Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	0.2574	0.0902	0.0807	0.4342	8.15	0.0043
post	1	-0.0648	0.1273	-0.3143	0.1847	0.26	0.6107
month12	1	0.1466	0.1277	-0.1037	0.3969	1.32	0.2509
ICgroup	1	0.0163	0.1276	-0.2338	0.2664	0.02	0.8985
post*ICgroup	1	0.3665	0.1818	0.0102	0.7227	4.07	0.0438
month12*ICgroup	1	-0.1205	0.1837	-0.4805	0.2395	0.43	0.5118
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

GEE Results for months 0, 6, 12 **AR(1)**

```
                         GEE Model Information

      Correlation Structure                              AR(1)
      Subject Effect                             id (1000 levels)
      Number of Clusters                                  1000
      Correlation Matrix Dimension                           3
      Maximum Cluster Size                                   3
      Minimum Cluster Size                                   3


Algorithm converged.


                    Working Correlation Matrix

                         Col1          Col2          Col3

        Row1           1.0000        0.3803        0.1446
        Row2           0.3803        1.0000        0.3803
        Row3           0.1446        0.3803        1.0000
```

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	0.2574	0.0902	0.0807	0.4342	2.85	0.0043
post	-0.0648	0.0985	-0.2580	0.1283	-0.66	0.5107
month12	0.1466	0.1036	-0.0564	0.3496	1.42	0.1568
ICgroup	0.0163	0.1276	-0.2338	0.2664	0.13	0.8985
post*ICgroup	0.3665	0.1444	0.0835	0.6495	2.54	0.0111
month12*ICgroup	-0.1205	0.1432	-0.4012	0.1603	-0.84	0.4003

GEE Results for months 0, 6, 12 **Unstructured**

GEE Model Information

Correlation Structure	Unstructured
Subject Effect	id (1000 levels)
Number of Clusters	1000
Correlation Matrix Dimension	3
Maximum Cluster Size	3
Minimum Cluster Size	3

Algorithm converged.

Working Correlation Matrix

	Col1	Col2	Col3
Row1	1.0000	0.3720	0.2737
Row2	0.3720	1.0000	0.3902
Row3	0.2737	0.3902	1.0000

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	0.2692	0.0896	0.0937	0.4448	3.01	0.0027
post	-0.0037	0.0906	-0.1812	0.1738	-0.04	0.9677
ICgroup	0.0065	0.1272	-0.2428	0.2559	0.05	0.9591
post*ICgroup	0.3163	0.1313	0.0589	0.5738	2.41	0.0160

group	month0	month6	month12
control	β_0	$\beta_0 + \beta_{\text{post}}$	$\beta_0 + \beta_{\text{post}}$
intervention	β_0 $+\beta_{\text{ICgroup}}$	$\beta_0 + \beta_{\text{post}}$ $+\beta_{\text{ICgroup}}$ $+\beta_{\text{ICgroup:post}}$	$\beta_0 + \beta_{\text{post}}$ $+\beta_{\text{ICgroup}}$ $+\beta_{\text{ICgroup:post}}$

GEE1 - testing hypotheses

Wald Tests

- $H_0: \quad \beta_j = 0$

$$\hat{\beta}_j / \widehat{\text{s.e.}} \sim N(0,1)$$

- $H_0: \quad \boldsymbol{\gamma} = 0$

$$\boldsymbol{\gamma} = (\beta_{j+1}, \beta_{j+2}, \ldots, \beta_{j+r})$$

$$\widehat{\boldsymbol{\gamma}}^T \boldsymbol{V}_{\gamma}^{-1} \widehat{\boldsymbol{\gamma}} \sim \chi^2(r)$$

$\boldsymbol{V}_{\gamma}$ is the empirical variance matrix corresponding to $\widehat{\gamma}$.

Summary

- GEE1 - focus on the marginal mean parameter $\boldsymbol{\beta}$.
- Flexible mean models.
- Choice of "working correlation models".
- Semiparametric since only first (and second) moment model(s).
- "sandwich estimator" for var$(\widehat{\boldsymbol{\beta}})$.
- Caveat: MCAR assumed.
- Caveat: time-dependent covariates and weighting.

- Note: Model versus Estimation versus Software

- Examples:

  HIVNET IC Analysis

  Madras Longitudinal Study of Schizophrenia

  (see chapter 11 of DHLZ)

  Progabide Seizure Count Data

The New England Journal of Medicine
10 SHATTUCK STREET, BOSTON, MASSACHUSETTS 02115-6094

Models may be inaccurate when assumptions are violated, important predictors are missing, data is missing and improper imputation methods used, or with overfitting. There are two concerns. The first is the use of GEE to estimate the 95% confidence intervals. This approach does not fully account for clustering. A more rigorous approach would have used multi level hierarchical modeling with MCMC simulation. This seems particularly important given the smaller sample size of the hospitals with lower volume. Although the authors state that random effects and three level models were performed with similar results, I suspect that the point estimates moved toward the mean and the confidence intervals widened perhaps sufficiently to include 1. No data is provided in the article.

Since explicit integration is avoided, the GEE methodology is definitely an important contribution to the estimation of models for longitudinal and clustered data. We use GEE for longitudinal data on respiratory infection in Section 9.2 where it is also compared to random effects modeling. Interestingly, GEE has recently been extended to factor models (Reboussin and Liang, 1998), where the dependence structure is of primary interest.

A rather severe limitation is that missing data can apparently only be handled under the restrictive assumption of missing completely at random MCAR (Liang and Zeger, 1986), since the estimating equations will otherwise be biased (e.g. Rotnitzky and Wypij, 1994). However, it is often not recognized that missingness may actually depend on covariates but not on observed responses (Little, 1995). Robins *et al.* (1994) suggest combining estimating equations with inverse probability weighting, yielding consistent estimators if the missing data mechanism is correctly specified.

Another limitation is that it is in general difficult to assess model adequacy in GEE (e.g. Albert, 1999); likelihood based diagnostics are for instance not available. The use of GEE should furthermore be reserved to problems where marginal or population averaged effects are of interest and avoided in analyses of etiology. This is because causal processes must operate at the cluster or individual level, not the population level. Population averaged effects are therefore merely descriptive and largely determined by the degree of heterogeneity in the population. Finally, Lindsey and Lambert (1998) and Crouchley and Davies (1999) point out that the estimated regression parameters are no longer consistent if there are endogenous covariates such as 'baseline' (initial) responses in longitudinal data.

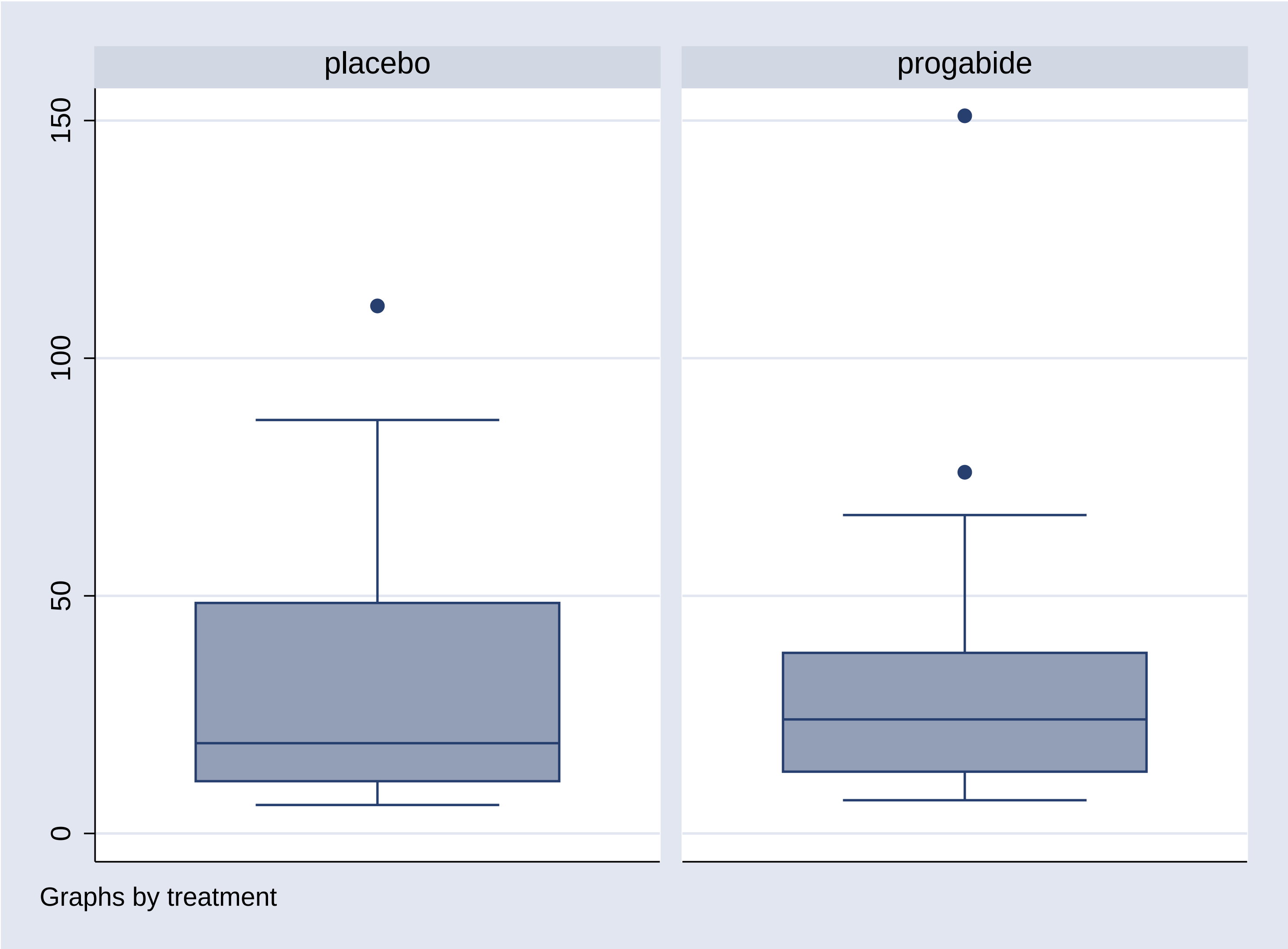
Example of Longitudinal Count Data

- Epileptic Seizures
  - ▷ **Subjects**: A total of N=59 patients were randomized to the anti-epileptic drug progabide, or to placebo in addition to standard chemotherapy.
  - ▷ **Baseline Measures**: Over an 8-week period prior to randomization a "baseline" number of seizures was recorded for each participant.
  - ▷ **Outcome**: Over (4) subsequent follow-up time periods the number of seizures in each 2-week period was recorded.
- **Q**: Is the drug progabide effective at reducing the rate of epileptic seizures?

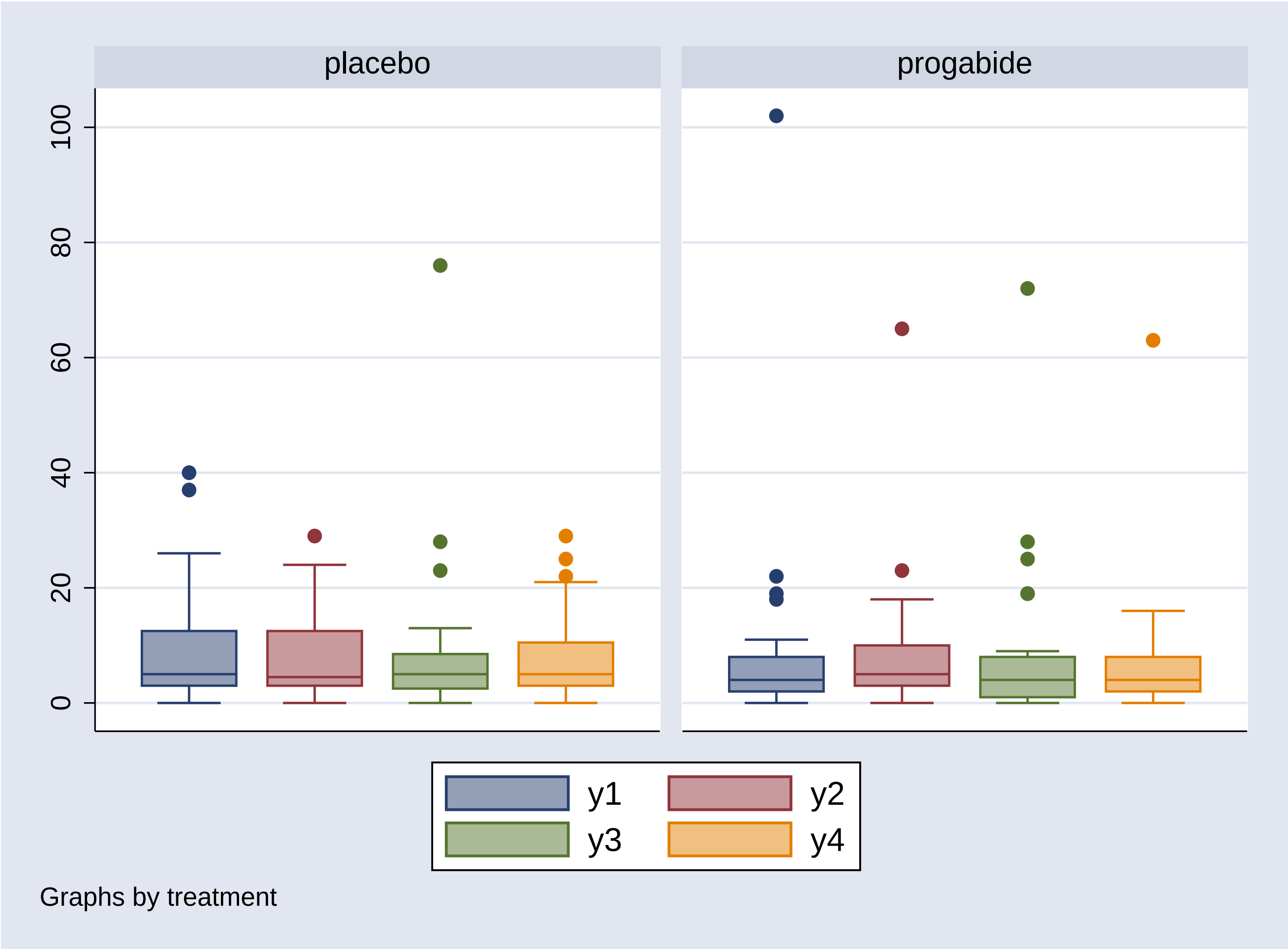
Analysis Options

- **Post-only** analysis using comparison of means, or Poisson regression.
  - ▷ Need to combine all post-baseline visits into single measurement, or choose a single (final, primary) outcome time.
- **Longitudinal** analysis.
  - ▷ Analysis of all data
  - ▷ Regression model for group and time
  - ▷ **Q**: How to model group and time?
  - ▷ **Q**: What will be the primary test for treatment differences?
    - ∗ At **any** time? (global test)
    - ∗ At **certain** time? (choose primary time)
  - ▷ **Q**: How to use baseline?

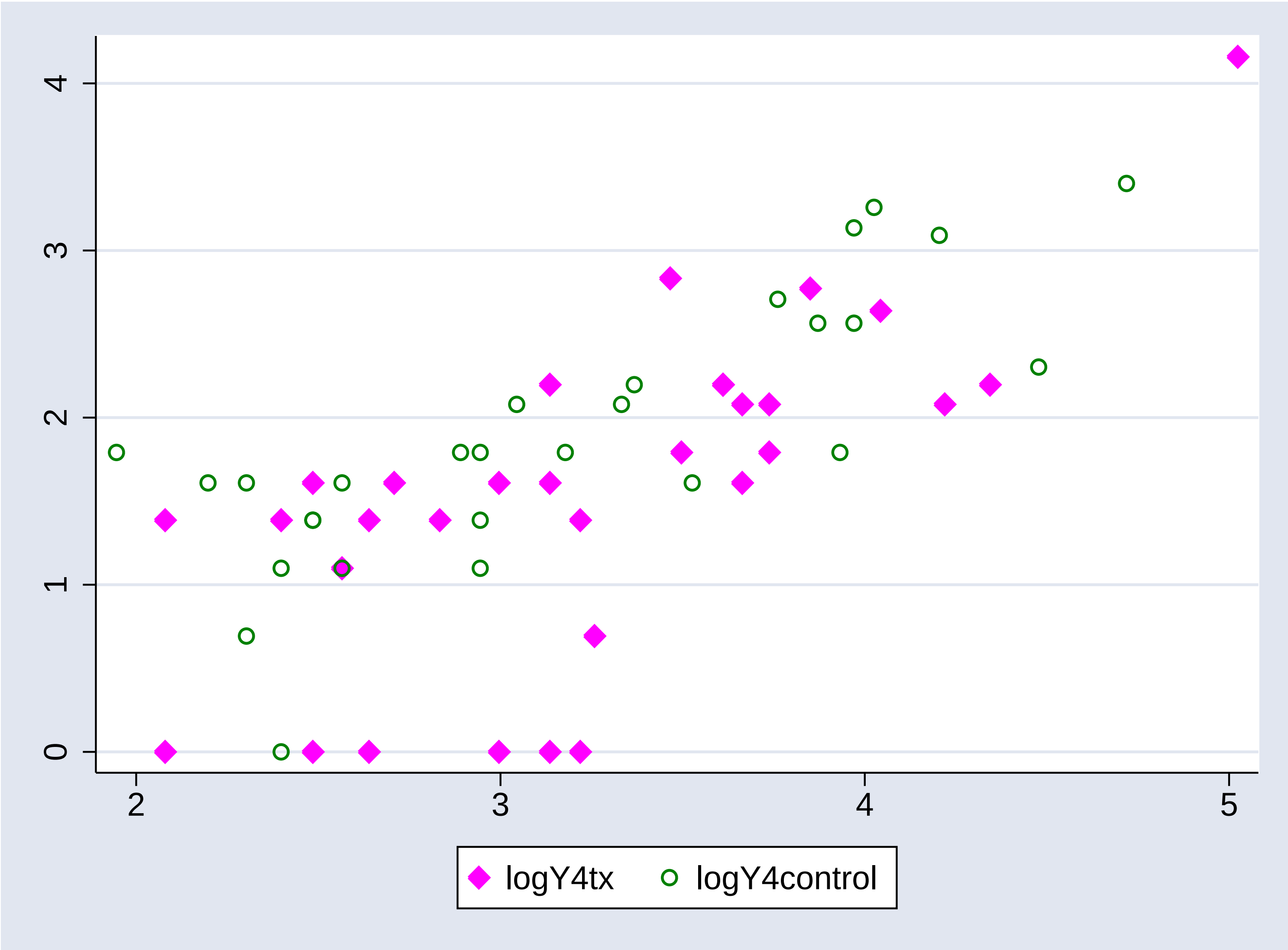
Seizure Data: Baseline (8 week period)



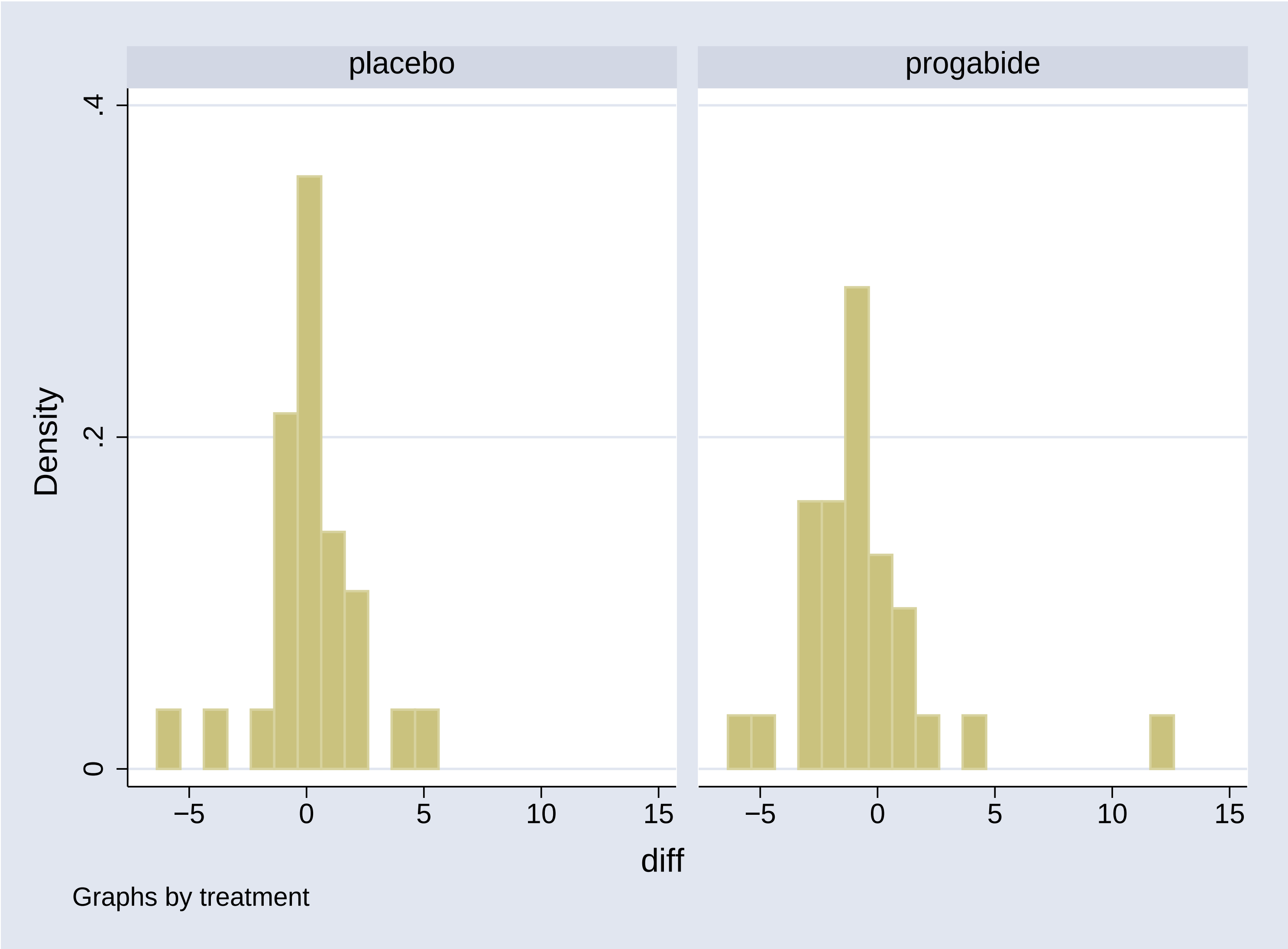
Seizure Data: Post Times (2 week periods)



Seizure Data: Post versus Pre



Seizure Data: Change (y4/2 - y0/8)



Seizure Data – Summaries

Variable	Obs	Mean	Std. Dev.	Min	Max
age	59	28.33898	6.301642	18	42

treatment	Freq.	Percent	Cum.
placebo	28	47.46	47.46
progabide	31	52.54	100.00
Total	59	100.00	

Seizure Data – Summaries

```
-> tx = placebo
Variable |       Obs        Mean    Std. Dev.       Min        Max
---------+-----------------------------------------------------
      y0 |        28    30.78571    26.10429          6        111
      y1 |        28    9.357143    10.13689          0         40
      y2 |        28    8.285714    8.164318          0         29
      y3 |        28    8.785714    14.67262          0         76
      y4 |        28    7.964286    7.627835          0         29

-> tx = progabide
Variable |       Obs        Mean    Std. Dev.       Min        Max
---------+-----------------------------------------------------
      y0 |        31     31.6129    27.98175          7        151
      y1 |        31    8.580645    18.24057          0        102
      y2 |        31    8.419355    11.85966          0         65
      y3 |        31    8.129032    13.89422          0         72
      y4 |        31    6.709677    11.26408          0         63
```

Seizure Data – Summaries

```
. *** CORRELATION exploratory analysis
-> tx = placebo (obs=28)
             |       y0       y1       y2       y3       y4
-------------+---------------------------------------------
          y0 |   1.0000
          y1 |   0.7442   1.0000
          y2 |   0.8313   0.7823   1.0000
          y3 |   0.4931   0.5070   0.6609   1.0000
          y4 |   0.8180   0.6746   0.7804   0.6757   1.0000
-----------------------------------------------------------------
-> tx = progabide  (obs=31)
             |       y0       y1       y2       y3       y4
-------------+---------------------------------------------
          y0 |   1.0000
          y1 |   0.8542   1.0000
          y2 |   0.8464   0.9070   1.0000
          y3 |   0.8350   0.9125   0.9249   1.0000
          y4 |   0.8750   0.9713   0.9466   0.9523   1.0000
-----------------------------------------------------------------
```

Regression Analysis

- $\boxed{\text{Poisson Regression}}$
  - **Outcome**: Y_{ij} seizure count at time t_{ij}
  - **Length of Observation**: T_j = 8 weeks, or 2 weeks
  - **Covariates**: $\texttt{Tx}_i$, t_{ij}.
- $\boxed{\text{Mean Model}}$

$$\mu_{ij} \quad = \quad \lambda_{ij} \cdot T_j \; = \texttt{Rate} \;\; \times \;\; \texttt{ObsTime}$$

$$\log \mu_{ij} \quad = \quad \underbrace{\beta_0 + \beta_1 \cdot t_{ij} + \beta_2 \cdot \texttt{Tx}_i + \beta_3 \cdot \texttt{Tx}_i \cdot t_{ij}}_{\log \lambda_{ij}} + \texttt{offset}(\log T_j)$$

STATA Analysis

```
*** LONGITUDINAL regression models

gen logY0 = ln( y0+1 )

save ThallWide, replace
reshape long y, i(id) j(week)

gen obsTime = 2*(week>0) + 8*(week==0)
gen logObsTime = log( obsTime )

*** create some variables
gen weekXtx = week * tx

*** GEE with all times as outcome
```

```
xtgee y week tx weekXtx, offset(logObsTime) ///
  i(id) corr(unstructured) t(week) family(poisson) link(log) robust

xtcorr

lincom tx + 4 * weekXtx
test tx weekXtx

*** DHLZ p. 165 Analysis of these data
gen post = (week>0)
gen postXtx = post * tx
xtgee y post tx postXtx, offset(logObsTime) ///
  i(id) corr(exchangeable) family(poisson) link(log) robust

xtcorr

lincom tx + postXtx
test tx postXtx
```

Seizure Analysis

```
. xtgee y week tx weekXtx, offset(logObsTime) ///
  i(id) corr(unstructured) t(week) family(poisson) link(log) robust

GEE population-averaged model
Group and time vars:                     id week
Link:                                        log
Family:                                  Poisson

                     (standard errors adjusted for clustering on id)
------------------------------------------------------------------------
           |           Semi-robust
         y |     Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------+------------------------------------------------------------
      week |   0.02131   .04230     0.50   0.614    -.06159    .10423
        tx |   0.01833   .22517     0.08   0.935    -.42300    .45967
   weekXtx |  -0.04117   .06673    -0.62   0.537    -.17197    .08961
     _cons |   1.32643   .16511     8.03   0.000    1.00281    1.6500
logObsTime |   (offset)
------------------------------------------------------------------------

.
```

```
. xtcorr

Estimated within-id correlation matrix R:

        c1      c2      c3      c4      c5
r1  1.0000
r2  0.9877  1.0000
r3  0.7106  0.8317  1.0000
r4  0.8008  0.9831  0.7326  1.0000
r5  0.6832  0.8089  0.5583  0.7112  1.0000

.
. lincom tx + 4 * weekXtx

 ( 1)  tx + 4 weekXtx = 0

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.1463748   .3672777    -0.40   0.690    -.8662259    .5734762
------------------------------------------------------------------------------
```

```
. test tx weekXtx

 ( 1)  tx = 0
 ( 2)  weekXtx = 0

           chi2(  2) =    0.40
         Prob > chi2 =    0.8176
```

Seizure Analysis

```
. *** DHLZ p. 165
. gen post = (week>0)
. gen postXtx = post * tx

. xtgee y post tx postXtx, offset(logObsTime) ///
   i(id) corr(exchangeable) family(poisson) link(log) robust
GEE population-averaged model
Link:                                        log
Family:                                  Poisson
Correlation:                        exchangeable
                     (standard errors adjusted for clustering on id)
------------------------------------------------------------------------
           |          Semi-robust
         y |    Coef.   Std. Err.      z     P>|z|    [95% Conf. Interval]
-----------+------------------------------------------------------------
      post |  0.11079   .11709      0.95   0.344    -.11870    .34030
        tx |  0.02651   .22375      0.12   0.906    -.41204    .46507
   postXtx | -0.10368   .21544     -0.48   0.630    -.52594    .31858
     _cons |  1.34760   .15870      8.49   0.000    1.03654   1.65867
logObsTime |   (offset)
------------------------------------------------------------------------
```

```
.
. xtcorr

Estimated within-id correlation matrix R:

        c1      c2      c3      c4      c5
r1  1.0000
r2  0.7769  1.0000
r3  0.7769  0.7769  1.0000
r4  0.7769  0.7769  0.7769  1.0000
r5  0.7769  0.7769  0.7769  0.7769  1.0000

.
. lincom tx + postXtx

 ( 1)  tx + postXtx = 0

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.0771661   .3570763    -0.22   0.829    -.7770228    .6226907
------------------------------------------------------------------------------
```

```
. test tx postXtx

 ( 1)  tx = 0
 ( 2)  postXtx = 0

           chi2(  2) =    0.31
         Prob > chi2 =    0.8543
```

STATA Analysis

```
*** GEE with BASELINE as covariate, and LINEAR model for time

xtgee y week tx weekXtx logY0 if week>0, offset(logObsTime) ///
  i(id) corr(unstructured) t(week) family(poisson) link(log) robust

xtcorr

lincom tx + 4* weekXtx
test tx weekXtx
```

Seizure Analysis

```
. xtgee y week tx weekXtx logY0 if week>0, offset(logObsTime) ///
   i(id) corr(unstructured) t(week) family(poisson) link(log) robust

GEE population-averaged model
Group and time vars:                 id week
Link:                                    log
Family:                              Poisson
Correlation:                    unstructured
                     (standard errors adjusted for clustering on id)
--------------------------------------------------------------------
           |          Semi-robust
         y |    Coef.   Std. Err.     z    P>|z|   [95% Conf. Interval]
-----------+--------------------------------------------------------
      week |  -0.04042   .06675    -0.61   0.545    -.17126    .09041
        tx |  -0.04387   .27064    -0.16   0.871    -.57433    .48658
   weekXtx |  -0.02914   .07721    -0.38   0.706    -.18048    .12218
     logY0 |   1.21558   .15635     7.77   0.000     .90913   1.52204
     _cons |  -2.72323   .63807    -4.27   0.000   -3.97384  -1.47262
logObsTime |   (offset)
----------------------------------------------------------------------
```

```
.
. xtcorr

Estimated within-id correlation matrix R:

        c1      c2      c3      c4
r1  1.0000
r2  0.4427  1.0000
r3  0.4270  0.5912  1.0000
r4  0.2674  0.2949  0.4427  1.0000
```

```
. lincom tx + 4* weekXtx

 ( 1)  tx + 4 weekXtx = 0

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.1604703   .2138171    -0.75   0.453    -.5795441    .2586034
------------------------------------------------------------------------------
```

```
.
. test tx weekXtx

 ( 1)  tx = 0
 ( 2)  weekXtx = 0

           chi2(  2) =    0.56
         Prob > chi2 =    0.7545
```

Summary of Seizure Analysis

- GEE: Poisson regression for counts
- GEE: Correlation model, robust standard errors
- Baseline
- Models for time and group
- Inference/testing for group
- **Q**: Enough clusters to trust the **robust** standard error?

GEE and Small Number of Clusters

- A number of investigations have shown that the robust standard error is too small when there are "few" clusters.
- Sharples and Breslow (1992); Emrich and Piedmonte (1992).
- With a small number of clusters the standard error is too small. This leads to tests (estimate/s.e.) that are larger than they should be and thus the null hypothesis is rejected more than the nominal 5% rate.
- Mancl and DeRouen (2001) present a simulation study of binary outcomes, with some suggested alternatives to the basic robust variance.
    - n=32 obs/cluster on average
    - intra-cluster correlation of 0.3

		Type 1 Error	
	cov (s.e.)	cluster	observation
clusters	estimator	covariate $(X_{1,i})$	covariate $(X_{2,ij})$
10	robust	0.139	0.154
	jackknife	0.114	0.112
20	robust	0.109	0.136
	jackknife	0.058	0.077
30	robust	0.088	0.089
	jackknife	0.058	0.054
40	robust	0.074	0.094
	jackknife	0.050	0.068

GEE and Small Number of Clusters

- An alternative estimate of the standard error based on the **jackknife** performs better.
  - ▷ The jackknife estimates the regression coefficient multiple times, where an estimate $\widehat{\boldsymbol{\beta}}_{(i)}$ is obtained with **subject i's** data left out.
  - ▷ A final variance (standard error) estimate is based on the variance of these jackknife estimates – with a rescaling of $(N-1)/N$ where N is the number of clusters.
  - ▷ STATA: `jknife` command!

STATA Analysis – jackknife

```
jknife "xtgee y post tx postXtx, offset(logObsTime) i(id) corr(exchangeable)
   family(poisson) link(log) robust" _b, cluster(id)

command:       xtgee y post tx postXtx , offset(logObsTime) i(id)
               corr(exchangeable) family(poisson) link(log) robust

statistics:    b_post     = _b[post]
               b_tx       = _b[tx]
               b_postXtx  = _b[postXtx]
               b_cons     = _b[_cons]
```

- NOTE: The option `_b` asks for the jackknife coefficient estimates to be saved and then summarized

STATA Analysis – jackknife

```
Variable          |      Obs    Statistic    Std. Err.    [95% Conf. Interval]
------------------+-----------------------------------------------------------
b_post            |
          overall |       59     .1107981
           jknife |              .1172237     .1258157   -.1346237    .3690712
b_tx              |
          overall |       59     .0265146
           jknife |              .0265906     .2354094   -.4446326    .4978137
b_postXtx         |
          overall |       59    -.1036807
           jknife |             -.0673245     .2530788   -.5739168    .4392677
b_cons            |
          overall |       59     1.347609
           jknife |              1.361116     .1656826    1.029466    1.692766
```

- Compare standard errors to those on p. 377.