Longitudinal analysis

CMED6040 – Session 6

Tim Tsang (matklab@hku.hk)

School of Public Health
The University of Hong Kong

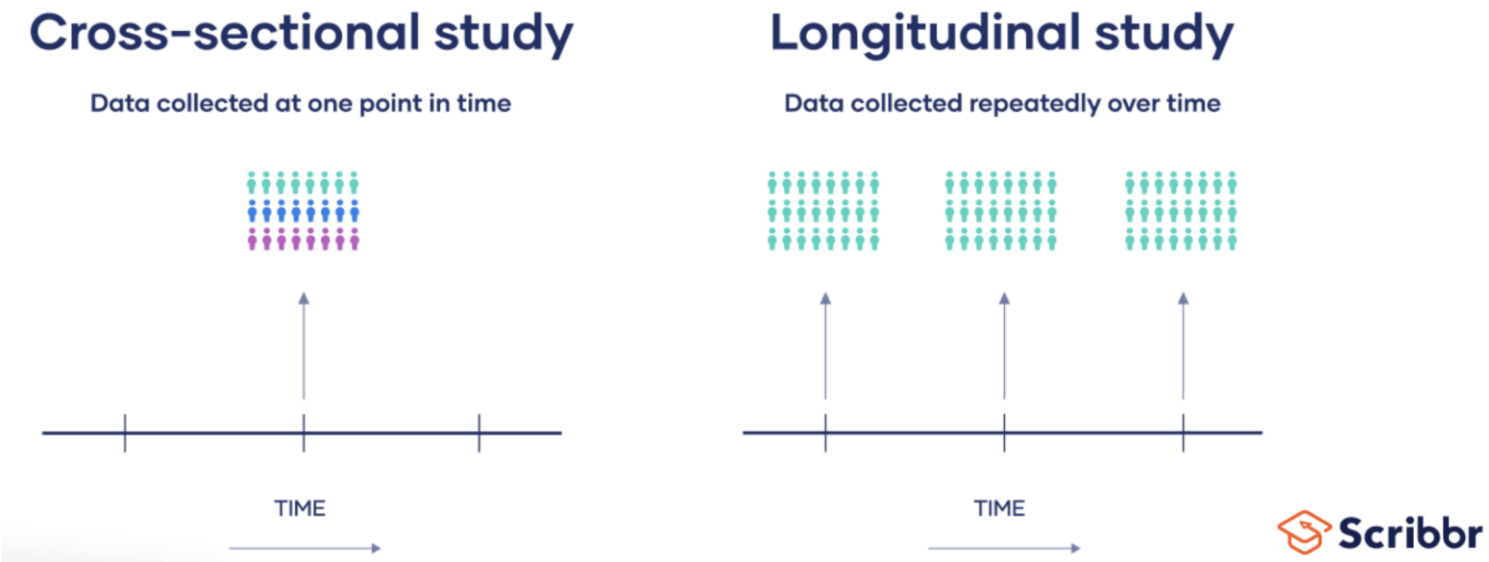
20 June 2023

Session 6 learning objectives

After this session, students should be able to

- Recognize and describe correlations between multiple measurements
- Analyse longitudinal data using generalized estimating equations (GEE)
- Perform model selection for variables and correlation structure for GEE

Longitudinal study

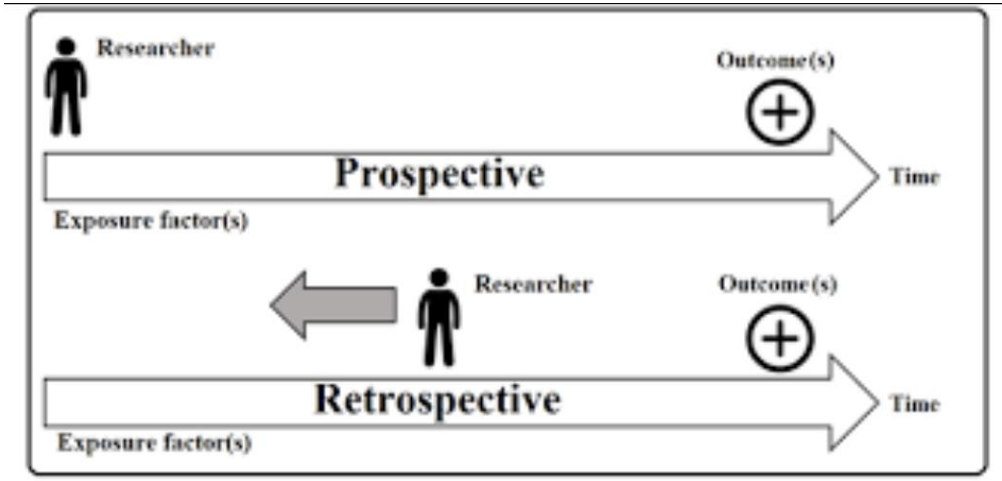


- Measurements were made repeatedly over time for each subject
- Provide stronger evidence on the causal effect
- Allow observation of change in subjects
- Higher power given the same number of subjects
- But more costly to carry out

Longitudinal data

- Data can be collected

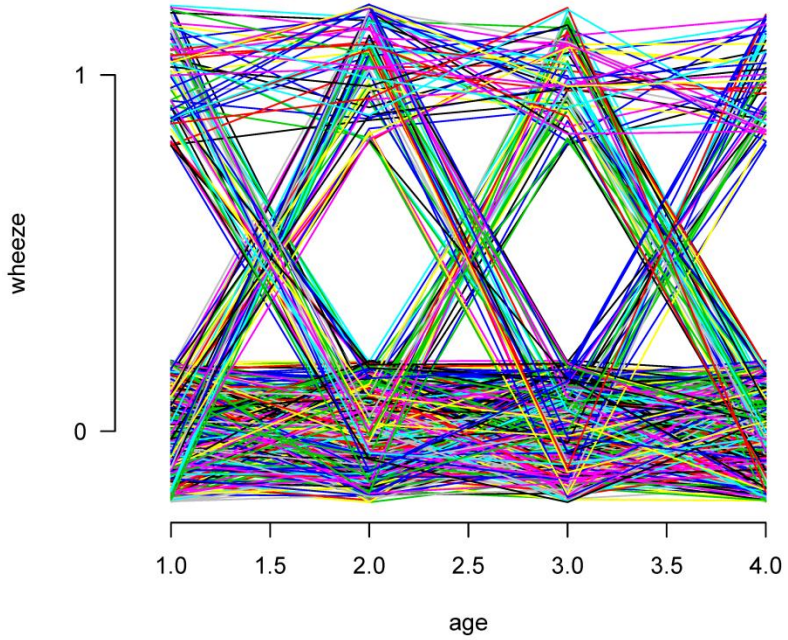
retrospectively or prospectively



- Subjects are usually assumed to be independent
- Within subject, the measurements are usually correlated
  - Statistical methods which account for such correlation are needed for correct inference
- Assuming a same relation across subjects, data from different subjects provide the basis for inference
- Also called ‘panel data’ for sociologists and economists

Examples of longitudinal data

- Monthly CD4 counts of HIV patients
- Daily viral shedding since symptom onset of MERS patients
- Weekly cognitive function of patients with Schizophrenia
- Relation between alcohol use and anxiety symptoms at different ages in a birth cohort
- Drinking and driving behaviours among adolescents over years



Generalized estimating equations (GEE)

- Marginal model – response depends on the covariates only
- Extension of GLM for correlated or clustered data
- Model specification (g is the link function):

$$g(\boldsymbol{\mu}_i) = \boldsymbol{x}_i{}'\boldsymbol{\beta}$$

- Based on the quasi-likelihood, the generalized estimating equation is:

$$U(\beta) = \sum_{i=1}^{N} \boldsymbol{D}_i{}'\boldsymbol{V}_i{}^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = 0$$

where $D_i = \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)$, $\boldsymbol{V}_i$ is the variance-covariance matrix of the repeated measurements, also determined by the chosen glm family

- Parameters are estimated by setting $U(\beta) = 0$
- Correlation structure regarded as nuisance parameter

Key assumptions of GEE

- Measurements are independent across subjects
- Measurements can be correlated within subjects
- The linear predictor $g(\boldsymbol{\mu}_i) = \boldsymbol{x}_i{}'\boldsymbol{\beta}$ is correctly specified

Correlation structure

- Specify how the observations are correlated within subjects
- Commonly used working correlation structure:
  - independence, exchangeable, AR(1), unstructured

- Independence: $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

- Exchangeable: $\begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$

Correlation structure

- First-order autoregressive/AR(1): $\begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}$
- Unstructured: $\begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix}$
- The number of unknown parameters are different and are estimated from data
  - e.g. AR(1): 1 unknown parameter, unstructured: 3 unknown parameters for 3 repeated measurements, (k^2-k)/2 =k(k-1)/2 in general

Characteristics of GEE

- The estimated parameters are efficient if the correlation is correctly specified
- The estimated parameters are still unbiased even with misspecification of the correlation structure
- However, the standard error will be less accurate
  - Can use a robust estimator for the standard error ('sandwich' estimator)
- Interpretation of the estimated parameters similar to GLM: depending on the chosen link function g.
- Handling of missing data
  - Estimates are valid under MCAR
  - Subjects with <k observations will still provide information on the correlation structure

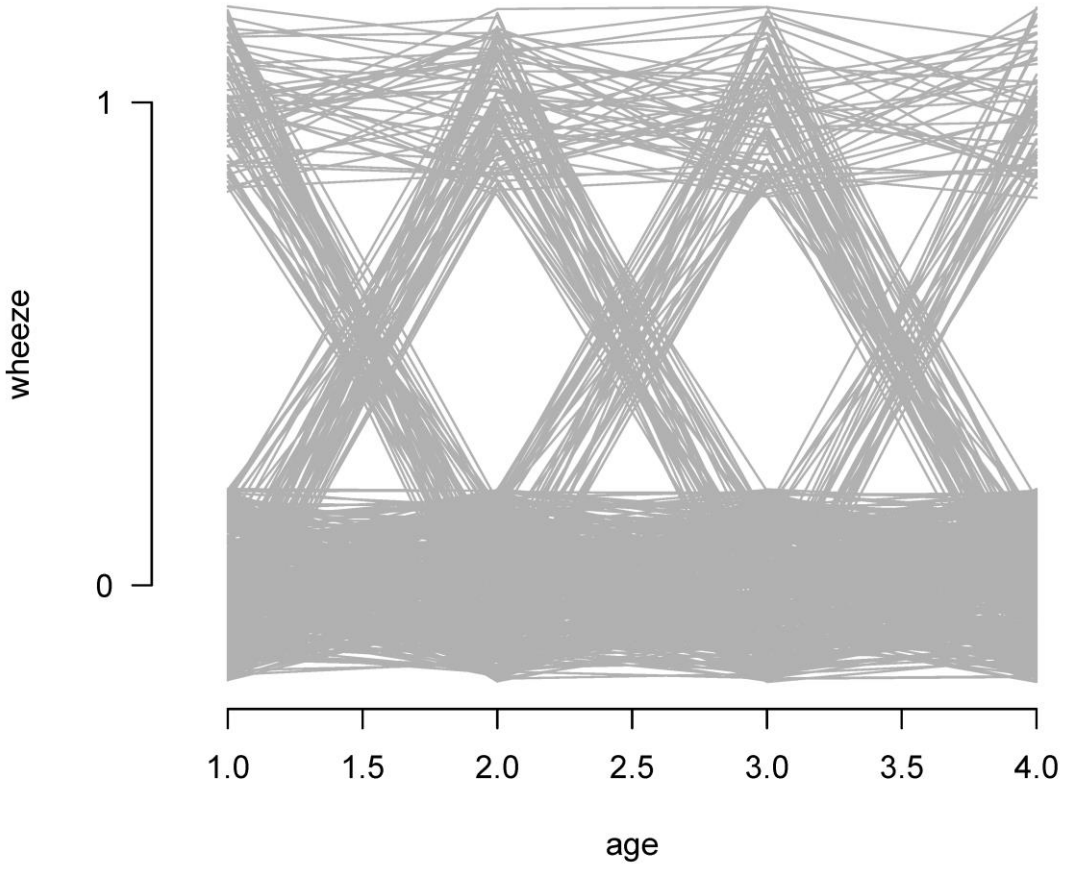
GEE in R

- package: geepack
- geeglm(*formula*, *family*, *corstr="independence"*, *id*, *data*, *subset*)
  - *formula, family, data, subset* same as glm()
  - *corstr* specifies the correlation structure, such as “independence” (by default), “exchangeable”, “ar1”, “unstructure”
  - *id* identifies the cluster/subject where multiple measurements were made

Example – health effect of air pollution

- Dataset from “geepack” package
- Can be loaded by data(ohio)
- Children were followed for four years, with wheeze status recorded annually
- Also information on age (0 = 9 years, time dependent) and maternal smoking status at the first year of the study (time independent)

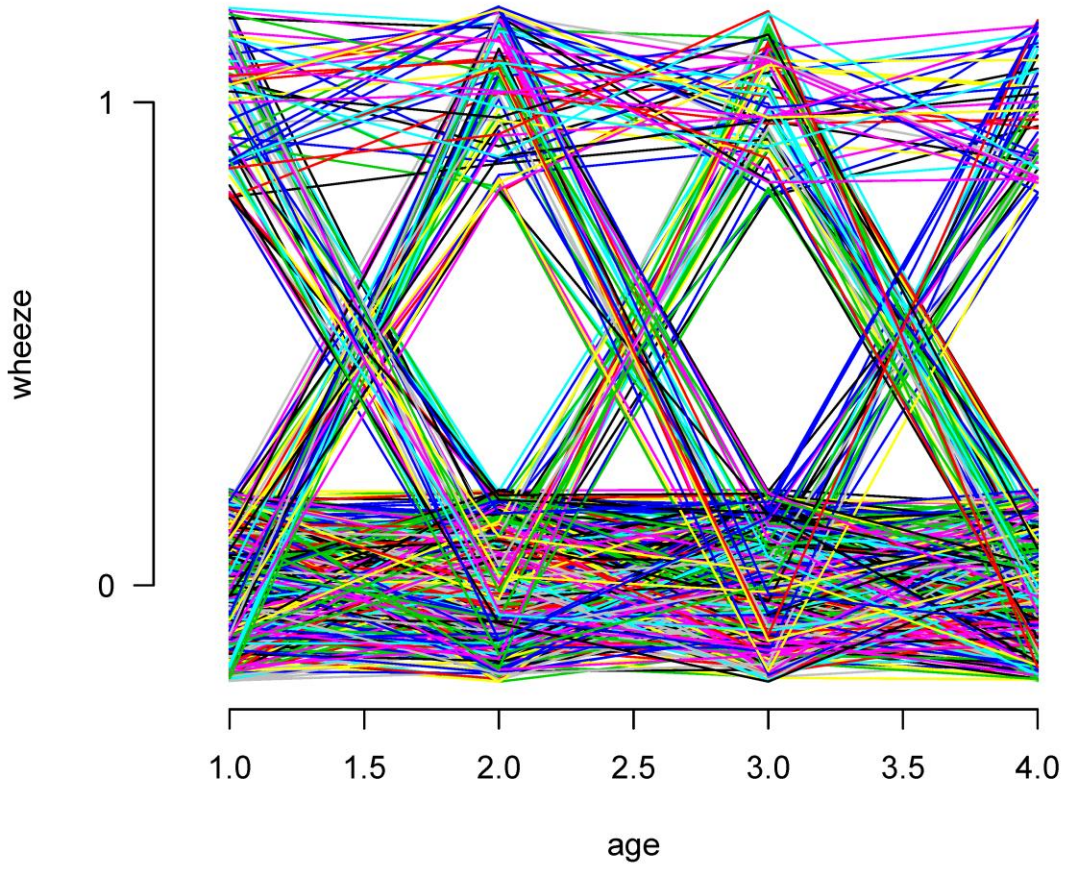
Plot data – ‘Spaghetti plot’




```
with(ohio,
interaction.plot(age, id,
jitter(resp), ylab='wheeze',
legend=F, lty=1,
col=gray(0.7)))
```

- to show temporal trends and corresponding proportions
- jitter() to avoid overlapping

Plot data – ‘Spaghetti plot’




```
with(ohio,
interaction.plot(age, id,
jitter(resp), ylab='wheeze',
legend=F, lty=1,
col=sample(1:20, max(id),
replace=T)))
```

Fitting a (naïve) GLM model

```
glm.ohio <- glm(resp~age+smoke, family=binomial, data=ohio)
summary(glm.ohio)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.8837     0.0838   -22.5   <2e-16 ***
age          -0.1134     0.0541    -2.1    0.036 *
smoke         0.2721     0.1235     2.2    0.028 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Age and maternal smoking are both significant at 5% sig. level

Correlation between the residuals

	7y	8y	9y	10y
7y	1	0.35	0.30	0.32
8y		1	0.44	0.33
9y			1	0.38
10y				1

- Observed within-subject correlation (also from the spaghetti plot)

```
cor(glm.ohio$residuals[ohio$age==A],glm.ohio$residuals[ohio$age==B])
```

Fitting a GEE model

- Suppose an independence correlation structure is assumed,

```
gee.indp <- geeglm(resp~age+smoke, family=binomial, data=ohio,
id=id, corstr = "independence")
summary(gee.indp)
Coefficients:
            Estimate Std.err   Wald Pr(>|W|)
(Intercept)  -1.8837  0.1142 271.90   <2e-16 ***
age          -0.1134  0.0439   6.68   0.0097 **
smoke         0.2721  0.1780   2.34   0.1263
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Same estimates as from GLM (but different standard errors)
- Standard error is the robust estimate (by default)
  - Still valid even with mis-specified correlation structure

Comparison between GLM and GEE

- GLM ignored the dependence between observations
- Comparison of the estimates (se):

	GLM	GEE (indept. corr.)
age	-0.11 (0.05)	-0.11 (0.04)
smoke	0.27 (0.12)	0.27 (0.18)

- Usually overestimate the standard errors of time-dependent predictors
  - Between-subject variability was not accounted for
- Usually underestimate the standard errors of time-independent predictors
  - Consider multiple measurements as additional independent samples

Fitting a GEE model

```
Coefficients:
             Estimate Std.err   Wald Pr(>|W|)
(Intercept)  -1.8837  0.1142 271.90   <2e-16 ***
age          -0.1134  0.0439   6.68   0.0097 **
smoke         0.2721  0.1780   2.34   0.1263
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Wald statistics: $\frac{(\hat{\theta}-\theta)^2}{var(\theta)} \sim \chi^2$

```
require(doBy)
esticon(gee.indp, c(0,0,1)) # specify a0β0+a1β1+a2β2+…
  beta0 Estimate Std.Error X2.value DF Pr(>|X^2|)   Lower Upper
1     0    0.272     0.178     2.34  1      0.126 -0.0767 0.621
```

Fitting a GEE model

- confint() doesn't work for geeglm object for obtaining confidence intervals
- Estimated ORs of the maternal smoking effect is 1.31 (95% CI = 0.93–1.86):

```
exp(c(esticon(gee.indp, c(0,0,1))$estimate,esticon(gee.indp,
c(0,0,1))$lwr,esticon(gee.indp, c(0,0,1))$upr))
  Estimate Lower Upper
1     1.31 0.926  1.86
```

- Estimated ORs of the age effect is 0.89 (95% CI = 0.82–0.97):

```
exp(c(esticon(gee.indp, c(0,1,0))$estimate,esticon(gee.indp,
c(0,1,0))$lwr,esticon(gee.indp, c(0,1,0))$upr))  Estimate Lower Upper
1    0.893 0.819 0.973
```

Fitting a GEE model

- On average, children with maternal smoking does not have a significantly different risk of wheezing
- On average, older children (between subjects) / children getting older (within subjects) will have a lower risk of wheezing

Model comparison

- Perform quasi-likelihood ratio test using anova() for nested models

```
gee.indp0 <- geeglm(resp ~ age, id=id, data=ohio,
family=binomial, corstr="independence")
anova(gee.indp, gee.indp0)

Analysis of 'Wald statistic' Table
Model 1 resp ~ age + smoke
Model 2 resp ~ age
  Df   X2 P(>|Chi|)
1  1 2.34      0.13
```

- Maternal smoking is not significant

Fitting GEE models with other correlation structure

- fit GEE models with exchangeable, AR(1) and unstructured correlation structure and compare the results

```
gee.exch <- geeglm(resp~age+smoke, family=binomial, data=ohio,
id=id, corstr = "exchangeable")
summary(gee.exch)
Coefficients:
            Estimate Std.err   Wald Pr(>|W|)
(Intercept)  -1.8804  0.1139 272.60   <2e-16 ***
age          -0.1134  0.0439   6.68   0.0097 **
smoke         0.2651  0.1777   2.22   0.1359
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.354  0.0624
```

Fitting GEE models with other correlation structure

- GEE model with AR(1) correlation structure

```
gee.ar1 <- geeglm(resp~age+smoke, family=binomial, data=ohio,
id=id, corstr = "ar1")

summary(gee.ar1)

 Coefficients:
             Estimate Std.err   Wald Pr(>|W|)
(Intercept)  -1.9022  0.1153 272.41   <2e-16 ***
age          -0.1149  0.0454   6.41    0.011 *
smoke         0.2345  0.1812   1.67    0.196
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
...
Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.491  0.0673
```

Fitting GEE models with other correlation structure

- GEE model with unstructured correlation structure

```
gee.unstr <- geeglm(resp~age+smoke, family=binomial, data=ohio,
id=id, corstr = "unstructured")

summary(gee.unstr)

 Coefficients:
             Estimate Std.err   Wald Pr(>|W|)
(Intercept)  -1.8886  0.1140 274.64   <2e-16 ***
age          -0.1149  0.0442   6.75   0.0094 **
smoke         0.2535  0.1782   2.02   0.1548
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fitting GEE models with other correlation structure

```
Estimated Correlation Parameters:
          Estimate Std.err
alpha.1:2    0.350  0.0732
alpha.1:3    0.308  0.0711
alpha.1:4    0.303  0.0710
alpha.2:3    0.470  0.0864
alpha.2:4    0.319  0.0736
alpha.3:4    0.376  0.0788
```

- Similar results for different correlation structure

Choosing a correlation structure

- Quasi-likelihood under the independence model information criterion (QIC)
- Model with a lower QIC is better
- Another correlation information criterion (CIC) was also proposed
- Available in the package “MESS”

```
require(MESS)
QIC(gee.indp); QIC(gee.exch); QIC(gee.ar1); QIC(gee.unstr)
           QIC    QICu     Quasi Lik  CIC params QICC
gee.indp  1829.5  1825.9   -909.9    4.80     3 1829.5
gee.exch  1829.5  1825.9   -909.9    4.80     3 1829.5
gee.ar1   1830.26 1826.27  -910.13   4.99     3 1830.27
gee.unstr 1829.60 1825.85  -909.97   4.82     3 1829.68
```

- All four models are similar
- QIC can also be used for selecting variables

Interaction between maternal smoking and age

```
gee.int.ar1 <- geeglm(resp~age*smoke, family=binomial, data=ohio,
id=id, corstr = "ar1")
summary(gee.int.ar1)
```

```
 Coefficients:
             Estimate Std.err   Wald Pr(>|W|)
(Intercept)  -1.9248  0.1207 254.31   <2e-16 ***
age          -0.1478  0.0598   6.10    0.014 *
smoke         0.2888  0.1914   2.28    0.131
age:smoke     0.0835  0.0917   0.83    0.362

Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.491   0.068
```

- No significant difference of the age effect across groups

Missing data

- Suppose some data were missing under MCAR:

```
set.seed(111)
n <- nrow(ohio)
n.missing <- 100
missing.x <- sample(1:n, n.missing, replace=F)
missing.y <- sample(c(3,4), n.missing, replace=T)

ohio.miss <- ohio
```

- Create a variable to specify the ordering of the repeated measures

```
ohio.miss$waves = ohio.miss$age + 3
ohio.miss[cbind(missing.x, missing.y)] <- NA
```

Missing data

- "waves" option in geeglm can specify the order of the observations
- Fit GEE with AR(1) correlation structure:

```
gee.ar1.miss <- geeglm(resp~age+smoke, family=binomial,
waves=waves, data=na.omit(ohio.miss), id=id, corstr = "ar1")
summary(gee.ar1.miss)
 Coefficients:
            Estimate Std.err   Wald Pr(>|W|)
(Intercept)   -1.890   0.117 263.24   <2e-16 ***
age           -0.101   0.047   4.58    0.032 *
smoke          0.188   0.184   1.04    0.307
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.499  0.0691
```

Model diagnostics

- GEE is an estimating procedure not based on formal likelihood function
- Standard likelihood-based goodness-of-fit statistics and model diagnostics not applicable
- Residual plots for assessing the specified mean model

GEE modelling strategy

- Modelling of the mean structure most important
- Use a reasonable correlation structure
  - Can assume independence for the working correlation structure if the expected correlation among repeated measurements is weak
  - Always use the robust estimate for the standard error
- Compare the estimates from different assumed correlation structure
- Can test different correlation structure by QIC

Analysis of longitudinal data using GLMM

- Fit the Ohio data using GLMM (in package lme4)

```
require(lme4)
ohio.glmm <- glmer(resp~age+smoke+(1|id), family=binomial,
data=ohio)
summary(ohio.glmm)
```

```
Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.374      0.275  -12.27   <2e-16 ***
age           -0.177      0.068   -2.60   0.0093 **
smoke          0.415      0.287    1.44   0.1485
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Similar results as from GEE
- GLMM estimates have a subject-specific interpretation

Review

- GEE estimates the population-averaged effects for longitudinal data
- Correlations among repeated measurements as nuisance parameters
- GEE is robust to mis-specification of the variance (or correlation structure)
- Model selection can be done by QIC or quasi-likelihood ratio test
- GLMM can also be used for longitudinal data, but the estimated effects have subject-specific interpretation

GEE example – children health insurance coverage

Research

Original Investigation

Effect of Expanding Medicaid for Parents on Children's Health Insurance Coverage

Lessons From the Oregon Experiment Randomized Trial

Jennifer E. DeVoe, MD, DPhil; Miguel Marino, PhD; Heather Angier, MPH; Jean P. O'Malley, MPH; Courtney Crawford, MPH; Christine Nelson, PhD, RN; Carrie J. Tillotson, MPH; Steffani R. Bailey, PhD; Charles Gallia, PhD; Rachel Gold, PhD, MPH

- DeVoe et al., JAMA Pediatr, 2015
- Objective: to estimate the effect on a child's health insurance coverage status when (1) a parent randomly gains access to health insurance and (2) a parent obtains coverage
- Subjects: 14,409 children in Oregon
- Outcome: Oregon Health Program (OHP) coverage (assessed monthly)

GEE example – children health insurance coverage

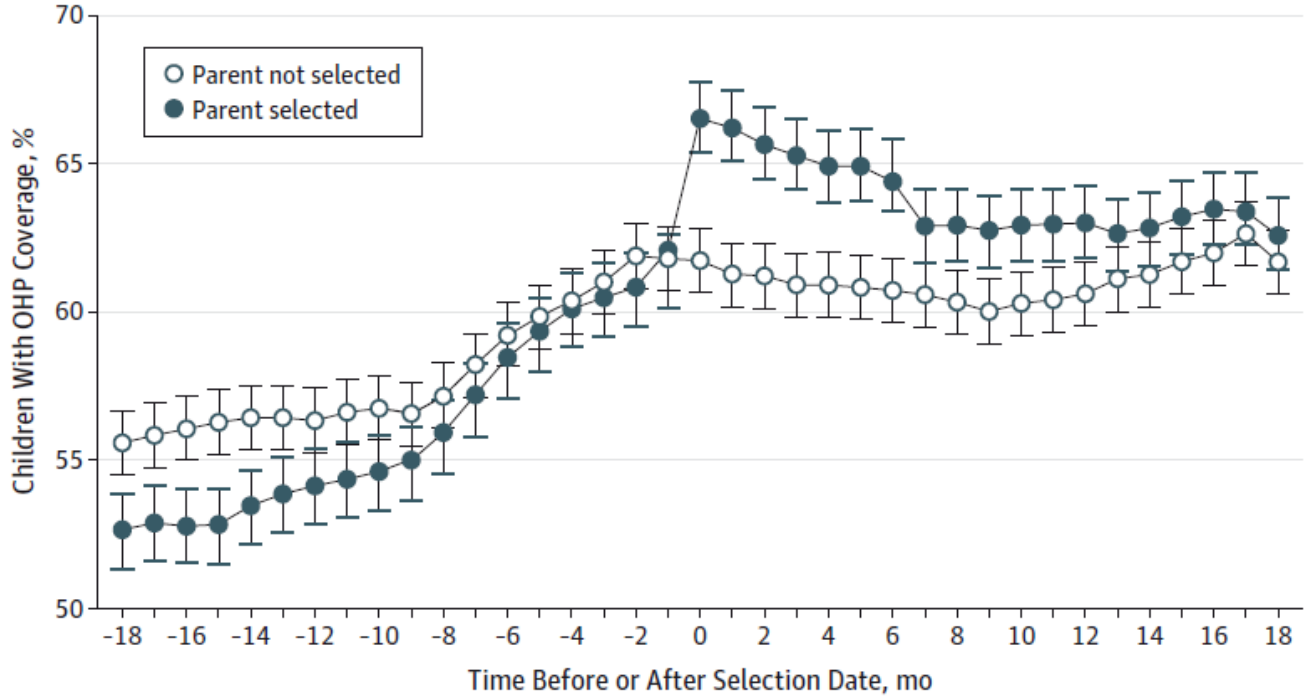
Statistical Analysis

We compared baseline characteristics between the selected and nonselected groups using Pearson χ^2 tests for categorical variables and Wilcoxon tests of differences for continuous variables. To examine the longitudinal effect of parental selection on child's insurance, we used a generalized estimating equation (GEE) model with a logit link and robust sandwich variance estimator to account for the temporal correlation of children's coverage during the study period. This model used child's insurance status in a given month as the outcome and was evaluated in each of the 18 months before and after the parental selection date. To estimate the effect parental selection status had on children's coverage after the selection date (intent-to-treat analyses), we used GEE models (as described earlier) limited to the 18 months after selection and summarizing the child's insurance for three 6-month intervals (0-6 months, 7-12 months, and 13-18 months after the parental selection date). We conducted per-protocol analyses using GEE models limited to children whose parents were selected and obtained OHP coverage (covered ≥50% of the time) in the first 6 months after the selection date (intervention group) and children whose parents were not selected and did not have OHP coverage in the first 6 months after selection (controls). In both the intent-to-treat and per-protocol models, we adjusted for covariates that significantly differed between the 2 groups at baseline. We report odds ratios (ORs) in this study, and these estimates do not approximate relative risk because coverage is not rare in this study population.[28]

- GEE was used to account for temporal correlation
- Logit link for binary outcome
- Robust sandwich variance estimator was used
- Assumed correlation structure not described
- Estimated ORs were reported

GEE example – children health insurance coverage

Figure 2. Percentage of Children With Oregon Health Plan (OHP) Coverage 18 Months Before and After Random Selection of Parents to Apply for OHP Coverage



GEE example – children health insurance coverage

Table 2. Effects of Parent Selection to Apply for OHP Coverage and of Selected Parent(s) Obtaining Coverage on Children Obtaining OHP Coverage

	Children With OHP Coverage by Time After Parent Selection[a]					
	1-6 mo		7-12 mo		13-18 mo	
Model	%[b]	OR (95% CI)	%[b]	OR (95% CI)	%[b]	OR (95% CI)
Intent-to-treat analyses[c]						
Unadjusted						
Parent selected[d]	65.8	1.26 (1.18-1.35)	63.1	1.19 (1.11-1.27)	63.2	1.14 (1.06-1.22)
Parent not selected[e]	60.5	1 [Reference]	59.0	1 [Reference]	60.2	1 [Reference]
Adjusted[f]						
Parent selected[d]	65.5	1.18 (1.10-1.27)	62.6	1.11 (1.03-1.19)	62.8	1.07 (0.99-1.14)
Parent not selected[e]	61.6	1 [Reference]	60.1	1 [Reference]	61.3	1 [Reference]
Per-protocol analyses[g]						
Unadjusted						
Parent selected and obtained OHP coverage[h]	76.5	2.55 (2.30-2.82)	70.6	1.92 (1.75-2.21)	69.4	1.67 (1.52-1.84)
Parent not selected and did not obtain OHP coverage[i]	56.0	1 [Reference]	55.5	1 [Reference]	57.6	1 [Reference]
Adjusted[f]						
Parent selected and obtained OHP coverage[h]	75.9	2.37 (2.14-2.64)	69.7	1.77 (1.60-1.96)	68.4	1.53 (1.38-1.69)
Parent not selected and did not obtain OHP coverage[i]	57.0	1 [Reference]	56.5	1 [Reference]	58.6	1 [Reference]

- Conclusions:
  - Children's odds of having OHP coverage increased when their parents were randomly selected to apply for Medicaid.
  - Children whose parents were selected and subsequently obtained coverage benefited most.

Conversion of data format in R (for reference)

- Can use reshape()
- Example – changing from 'long' form to 'wide' form:

```
data(ohio)
ohio$ex.age <- ohio$age+9
ohio.w <- reshape(ohio, v.names = c("age","resp"), idvar="id",
timevar = "ex.age", direction = "wide")
head(ohio.w)
```

```
   id smoke age.7 resp.7 age.8 resp.8 age.9 resp.9 age.10 resp.10
1   0     0    -2      0    -1      0     0      0      1       0
5   1     0    -2      0    -1      0     0      0      1       0
9   2     0    -2      0    -1      0     0      0      1       0
13  3     0    -2      0    -1      0     0      0      1       0
17  4     0    -2      0    -1      0     0      0      1       0
21  5     0    -2      0    -1      0     0      0      1       0
```

Conversion of data format in R (for reference)

- Example – changing from ‘wide’ form to ‘long’ form:

```
ohio.l <- reshape(ohio.w, varying = list(c(3,5,7,9),c(4,6,8,10)),
v.names = c("age","resp"), idvar="id", times=1:4, direction =
"long")
ohio.l <- ohio.l[order(ohio.l$id, ohio.l$time),]
head(ohio.l)
```

```
    id smoke time age resp
0.1  0     0    1  -2    0
0.2  0     0    2  -1    0
0.3  0     0    3   0    0
0.4  0     0    4   1    0
1.1  1     0    1  -2    0
1.2  1     0    2  -1    0
```

References

- Twisk JWR. Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide. Cambridge University Press, 2013.

- Diggle P, Heagerty P, Liang KY and Zeger SL. Analysis of Longitudinal Data, Oxford University Press, 2013.