Advanced GLMs: Analysis of Correlated Data

- Patrick J. Heagerty PhD
- Department of Biostatistics
- University of Washington

Bio & Notes

- **Patrick J. Heagerty**
  - Professor, University of Washington
  - Collaborative roles = CBS, VA ERIC, NIAMS MCRC, KL2
  - Books:

    Diggle, Heagerty, Liang & Zeger
    "Analysis of Longitudinal Data" Oxford, 2002.

    van Belle, Fisher, Heagerty & Lumley
    "Biostatistics" Wiley, 2004.
    (introductory chapter on LDA)

- **Course Notes & Slides**
  - UW Biostat 571 = Ph.D. applied core sequence
    Winter 2000, 2001, 2002, 2003, 2007
  - UM Epi 766 = Longitudinal Data Analysis / Epi
    Summer 2000 (Summer 2004 with VA/UW Biostat/Epi)
  - Second Seattle Symposium (with S. Zeger)
    Fall 2000
  - RAND short course; NICHD short course
    Fall 2002; Fall 2003
  - UW Biostat 540 = M.S. applied core sequence
    Spring 2005, 2006, 2007, 2008

Introduction

Objectives:

- Appreciate breadth of applications.

- Understand that correlation interacts with covariate design to impact standard errors.

- Understand that variance/covariance model is useful for efficiency of estimation.

Biostat 571 – Overview

$\boxed{\star}$ We will study methods (ie. theory & practice) for data with non i.i.d. errors:

Part I – **Generalized linear models** $approx$ (2 weeks)

◦ Review independent data with non-constant variance.

◦ Extend linear model by

- replace linear model for $\mu = E(Y)$ by linear model for $g(\mu)$.
- replace constant variance assumption with mean-variance relationship.
- replace normal distribution with exponential family.

- Models for multinomial outcomes (ie. the simplest "multivariate" response).

- Models / methods for "extra variation" = overdispersion.

Motivation

- Coronary artery disease (CAD) is the leading cause of death in men and women in the US.
- The "reference test" for CAD diagnosis is coronary contrast angiography. This test is invasive.
- "Stress" tests are a common method used for CAD diagnosis. This involves stimulation of the heart and imaging of the heart.

Stimulation	=	exercise,
		pharmacologic stressors
Imaging	=	echocardiography (ECHO),
		single photon emission computed
		tomography (SPECT)

Meta-analysis

- Many studies have investigated the accuracy of stress tests for the diagnosis of CAD.
- Systematic Reviews of Diagnostic Accuracy
  - Cochrane Methods Group – provides guidelines.
  - Goals include:
    1. Provide an overall summary of diagnostic accuracy (sensitivity, specificity).
    2. Compare different tests.
    3. Characterize **systematic** variation in accuracy (ie. subgroups of patients defined by gender, age, ...).
    4. Characterize **random** study-to-study variation.

Data

- Data extracted for (2) pharmacologic stressors:
  - ▷ Dobutamine: increases myocardial demand by increasing heart rate and contractility (like exercise)
  - ▷ Persantine: vasodilator of the epicardial coronary arteries. Leads to a "steal" of blood flow away from diseased areas.
- We have combined ECHO and SPECT imaging for plots.
- Data:
  - ▷ Sensitivity, specificity, and covariates from study i.
  - ▷ (Y_{i1}, N_{i1}), (Y_{i0}, N_{i0}), and $\boldsymbol{X}_i$.

    $N_{i1} = \#$ of diseased subjects in study i.

    $Y_{i1} = \#$ of diseased subjects that test positive.

    $N_{i0} = \#$ of non-diseased subjects in study i.

    $Y_{i0} = \#$ of non-diseased subjects that test positive.

Diagnostic Accuracy

- Consider a single cross sectional sample, a binary test, and a binary disease variable.

	$T+$	$T-$	
D	n_{11}	n_{10}	n_D
$\overline{D}$	n_{01}	n_{00}	$n_{\overline{D}}$
	n_{T+}	n_{T-}	N

Diagnostic Accuracy

Predictive probabilities:

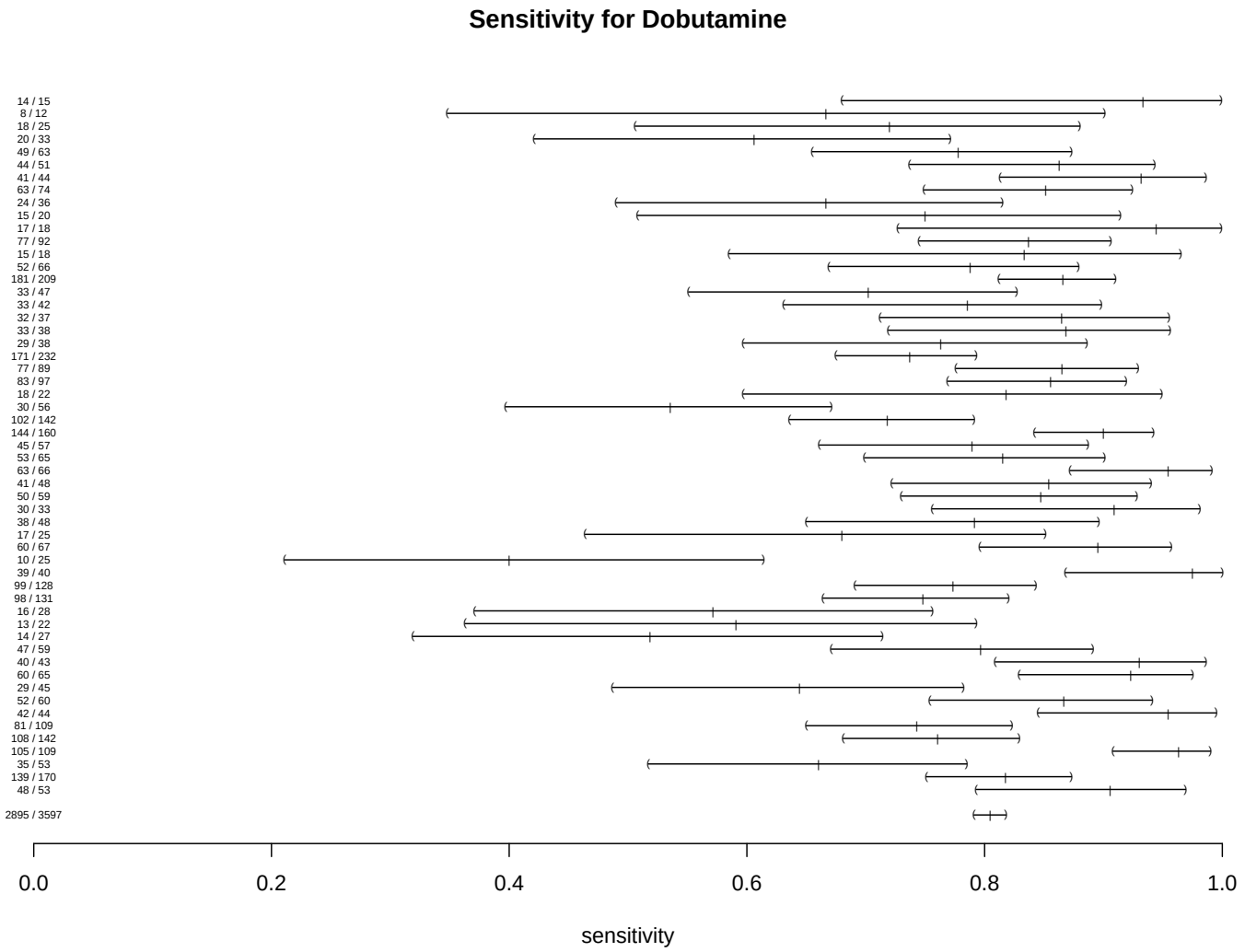
$$P[D \mid T+]$$

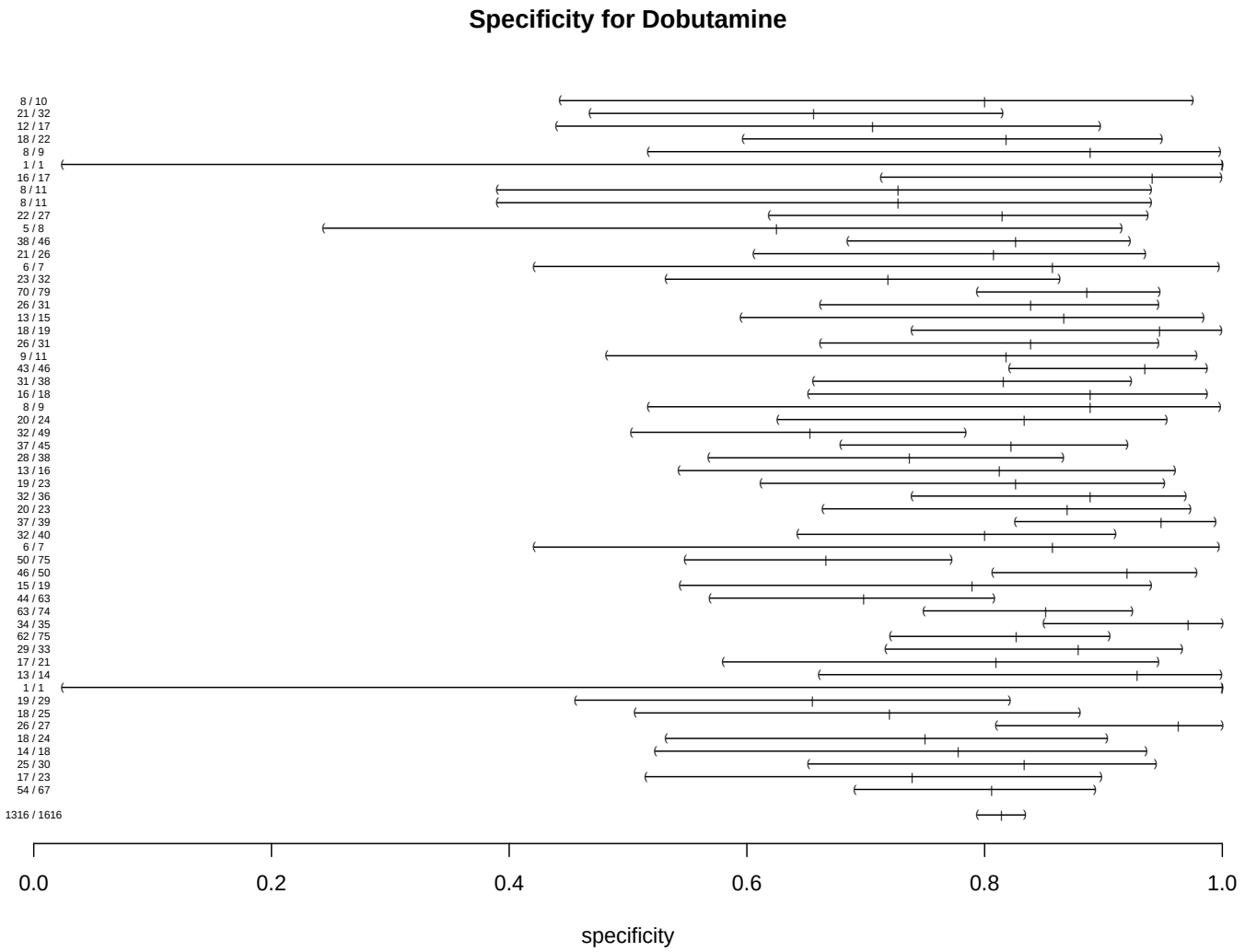
$$P[\overline{D} \mid T-]$$

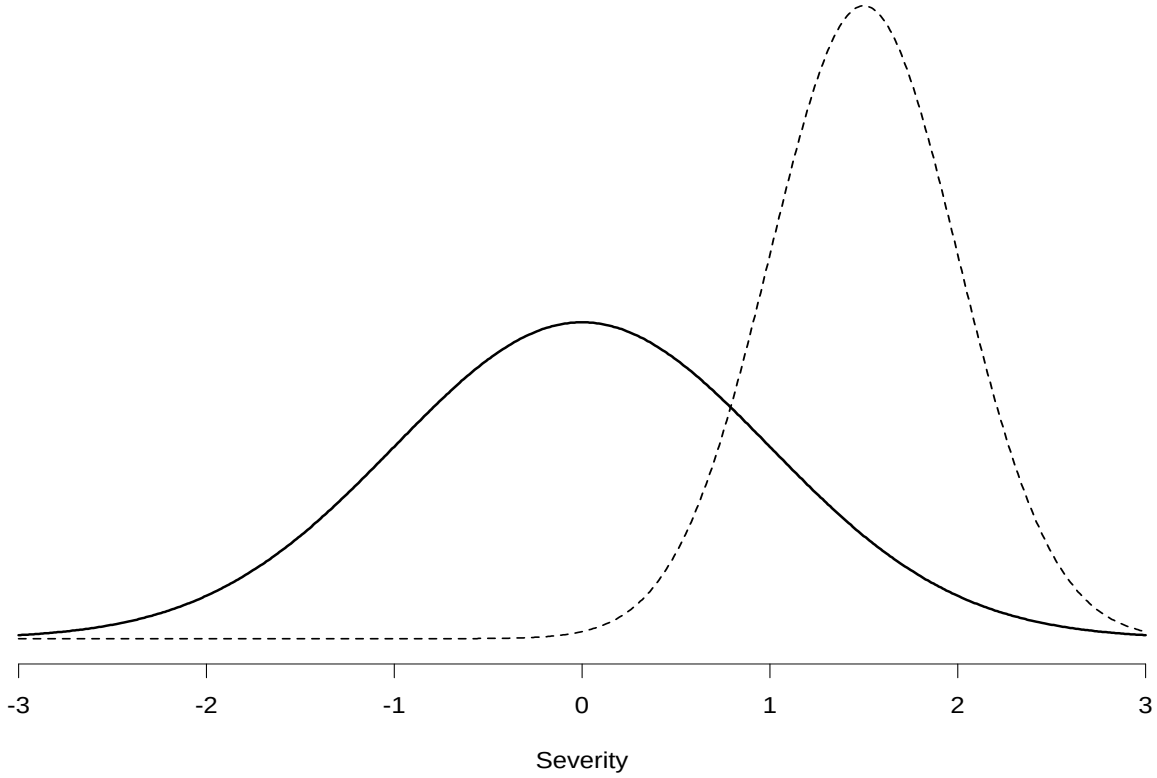
Accuracy summaries:

$$P[T+ \mid D]$$

$$P[T- \mid \overline{D}]$$


Sensitivity for Dobutamine
14 / 15
8 / 12
18 / 25
20 / 33
49 / 63
44 / 51
41 / 44
63 / 74
24 / 36
15 / 20
17 / 18
77 / 92
15 / 18
52 / 66
181 / 209
33 / 47
33 / 42
32 / 37
33 / 38
29 / 38
171 / 232
77 / 89
83 / 97
18 / 22
30 / 56
102 / 142
144 / 160
45 / 57
53 / 65
63 / 66
41 / 48
50 / 59
30 / 33
38 / 48
17 / 25
60 / 67
10 / 25
39 / 40
99 / 128
98 / 131
16 / 28
13 / 22
14 / 27
47 / 59
40 / 43
60 / 65
29 / 45
52 / 60
42 / 44
81 / 109
108 / 142
105 / 109
35 / 53
139 / 170
48 / 53
2895 / 3597
0.0
0.2
0.4
0.6
0.8
1.0
sensitivity


Specificity for Dobutamine
8 / 10
21 / 32
12 / 17
18 / 22
8 / 9
1 / 1
16 / 17
8 / 11
8 / 11
22 / 27
5 / 8
38 / 46
21 / 26
6 / 7
23 / 32
70 / 79
26 / 31
13 / 15
18 / 19
26 / 31
9 / 11
43 / 46
31 / 38
16 / 18
8 / 9
20 / 24
32 / 49
37 / 45
28 / 38
13 / 16
19 / 23
32 / 36
20 / 23
37 / 39
32 / 40
6 / 7
50 / 75
46 / 50
15 / 19
44 / 63
63 / 74
34 / 35
62 / 75
29 / 33
17 / 21
13 / 14
1 / 1
19 / 29
18 / 25
26 / 27
18 / 24
14 / 18
25 / 30
17 / 23
54 / 67
1316 / 1616
0.0
0.2
0.4
0.6
0.8
1.0
specificity



$\overline{D}$ D

- Define a positive test: $T+ = \mathbf{1}(Y > c)$.
- Two error rates for decisions.
- Test "makers" and test "takers".

Accuracy Summaries

- Sensitivity:

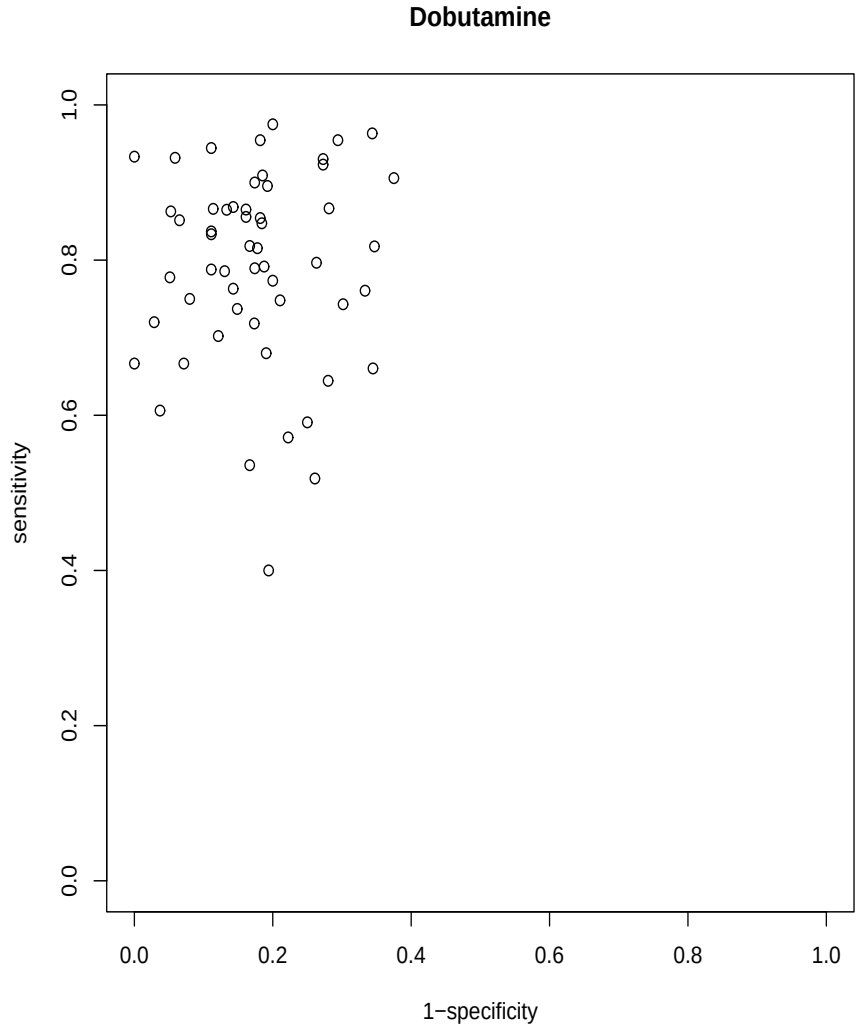
$$P[\text{ Test Positive } | \text{ Diseased }]$$

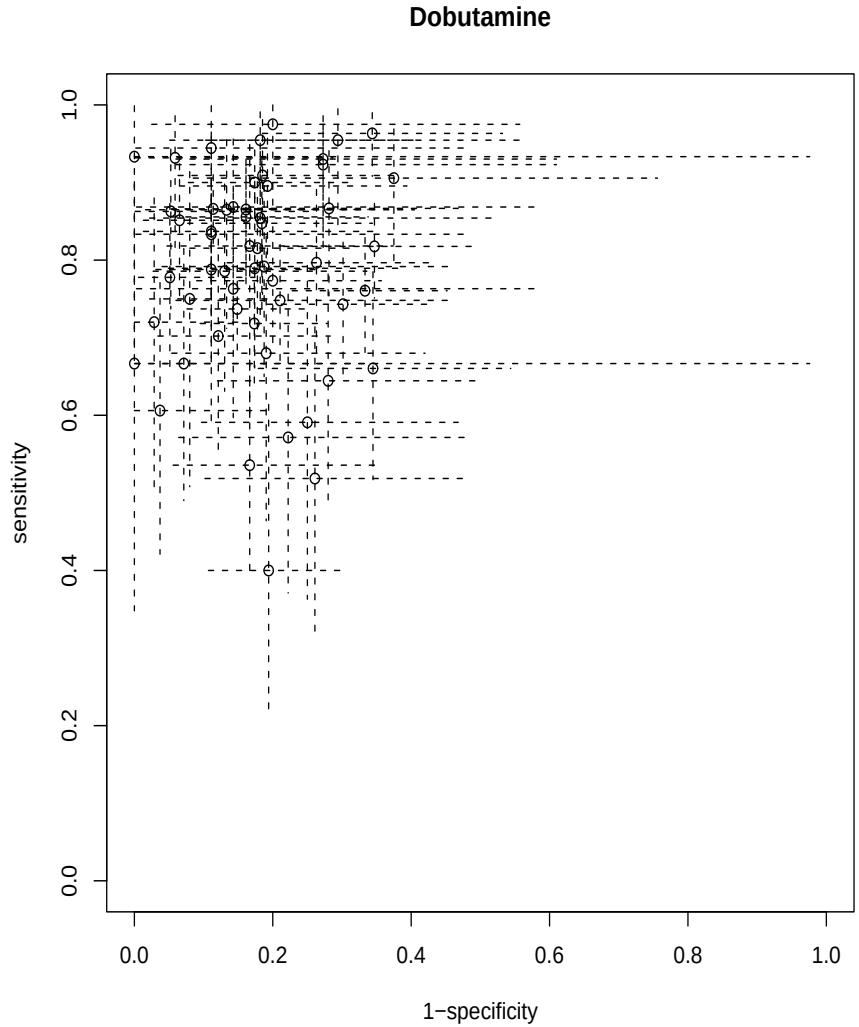
- Specificity:

$$P[\text{ Test Negative } | \text{ non-Diseased }]$$

$$1 - P[\text{ Test Positive } | \text{ non-Diseased }]$$

- ROC Curve:
  - Used when a positive test is defined by $Z > c$ for a continuous test, Z, and a "threshold" value, c.
  - points $[\ FP(c),\ TP(c)\ ]\ \ \forall\, c \in (-\infty, +\infty)$


Dobutamine
sensitivity
1−specificity
0.0
0.2
0.4
0.6
0.8
1.0


Dobutamine
sensitivity
1−specificity
0.0
0.2
0.4
0.6
0.8
1.0

Using Generalized Linear Models

- Compare the test modalities (echo, spect) $\times$ (dob, per).
- Analysis of sensitivity using binomial logistic regression.

```
#
# cad.roc.regn.q
#
# ------------------------------------------------------------
#
# PURPOSE:  run regression for the CAD data.
#
# DATE:  00/10/25
#
# AUTHOR:  P. Heagerty
#
#=====================================
#
# Variables:  (In column order of appearance)
#
# Y1            number of true-positive tests
# N1            number of diseased subjects
```

```
# DOBUTAMINE    1 if stimulant was dobutamine; 0 if persantine
# ECHO          1 if image modality was echo; 0 if spect
# YEAR          year of the study (minus 1999)
# AGE           average age in the study (minus 50)
# VERIFY        1 if no verification differential;
#                 0 if verification (bias)
# QUALITY       1 =  low quality; 2 = medium quality; 3 = high quality
# DEF50         1 = use of 50% stenoisis for CAD definition; 0 = use of 75%
# PERCAD        percent of study population with CAD
#
#=======================================
#
data <- read.table("cad.roc.data")
#
cad.data <- data.frame(
                          y = data[,1],
                          n = data[,2],
                          dob = data[,3],
                          echo = data[,4] )
#
fit0 <- glm( cbind( y, n-y ) ~ dob * echo,
                                family=binomial,
                                data=cad.data )
summary( fit0, cor=F )
#
```

```
fit1 <- glm( cbind( y, n-y ) ~ dob * echo,
                               family=quasi(
                                           link="logit",
                                           variance="mu(1-mu)" ),
                               data=cad.data )
summary( fit1, cor=F )
#
#
# end-of-file...
```

Binomial Regression Analysis

```
Call: glm(formula = cbind(y, n - y) ~ dob * echo,
                          family = binomial, data = cad.data)

Coefficients:
                Value Std. Error    t value
(Intercept)  2.1682444 0.07203410  30.100250
        dob -0.6595936 0.12349252  -5.341162
       echo -1.2863032 0.09082285 -14.162771
   dob:echo  1.1734438 0.14343752   8.180871

(Dispersion Parameter for Binomial family taken to be 1 )
    Null Deviance: 616.8647 on 114 degrees of freedom
Residual Deviance: 398.8679 on 111 degrees of freedom
```

Quasilikelihood Regression Analysis

```
Call: glm(formula = cbind(y, n - y) ~ dob * echo,
        family = quasi(link = "logit",
        variance = "mu(1-mu)"), data = cad.data)

Coefficients:
                 Value Std. Error    t value
(Intercept)  2.1682444  0.1383072  15.677013
        dob -0.6595936  0.2371087  -2.781820
       echo -1.2863032  0.1743821  -7.376349
   dob:echo  1.1734438  0.2754036   4.260816

(Dispersion Parameter for Quasi-likelihood family taken to be 3.686494 )
    Null Deviance: 616.8647 on 114 degrees of freedom
Residual Deviance: 398.8679 on 111 degrees of freedom
```

Statistical Issues

- Which (if any) of these analyses is valid?
  - ▷ **A**:
- How to interpret the resulting parameter estimates?
  - ▷ **A**:
- Are there other statistical approaches that may be more "appropriate"?
  - ▷ **A**:
- How to summarize the components of variability?
  - ▷ **A**:
- Should we jointly consider sensitivity and specificity?
  - ▷ **A**:

Biostat 571 – Overview

Part II – General LM for Correlated Continuous Data

approximately (4 weeks)

◦ Extend the linear model by considering a covariance structure for response vectors.

- Longitudinal data (repeated measures)
- Clustered data
- Multivariate response (MANOVA)
- Time-series and spatial data

Biostat 571 – Overview

- Semi-parametric methods
  - Weighted least squares
  - Empirical ("sandwich") variance estimates & efficiency
  - Specification and estimation of covariances
  - Inference

- Classical methods (ANOVA techniques).

- Methods based on multivariate Gaussian
  - Maximum likelihood (ML) and restricted ML (REML)
  - Linear mixed models
  - Prediction of random effects (empirical Bayes)
  - Longitudinal data analysis
  - Model checking (diagnostics)

Beta-carotene Phase II Data

Motivation:

- Beta-carotene is (was?) one of the most commonly used compounds in clinical trials of chemopreventive agents for various cancers.
- In 1992 a phase II study was conducted to examine the pharmokinetics of long-term, high-dose beta-carotene regimens.
- Interest is in the long-term dynamics of beta-carotene and the impact on alpha-tocopherol (vitamin E).

Beta-carotene Phase II Data

- Several time aspects are of interest:
  1. How long before stable plasma levels are obtained?
  2. Is the time course different depending on the dose of beta-carotene?
  3. Do changes in beta-carotene correlate with changes in vitamin E?

Data:

- The response variables are plasma concentration of beta-carotene and vitamine E.
- A total of 46 subjects were measured monthly for 3 months prior to randomization. Subjects were randomized to placebo, 15, 30, 45, or 60 mg/day for 9 months. Subjects were followed for an additional 3 months.
- Baseline patient factors include:

  `AGE` - age at randomization

  `MALE`

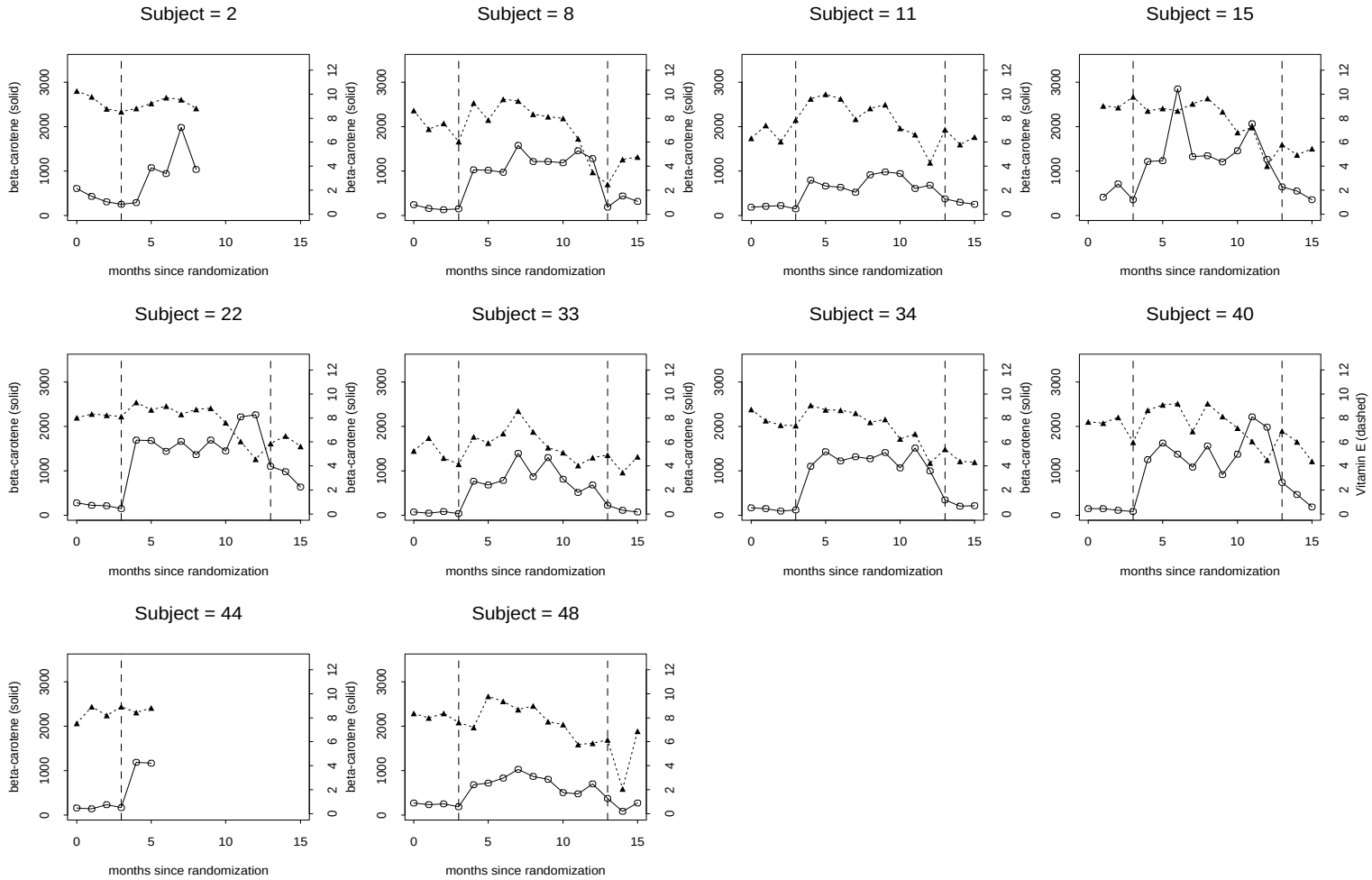
  `WEIGHT`

  `BMI` - body mass index

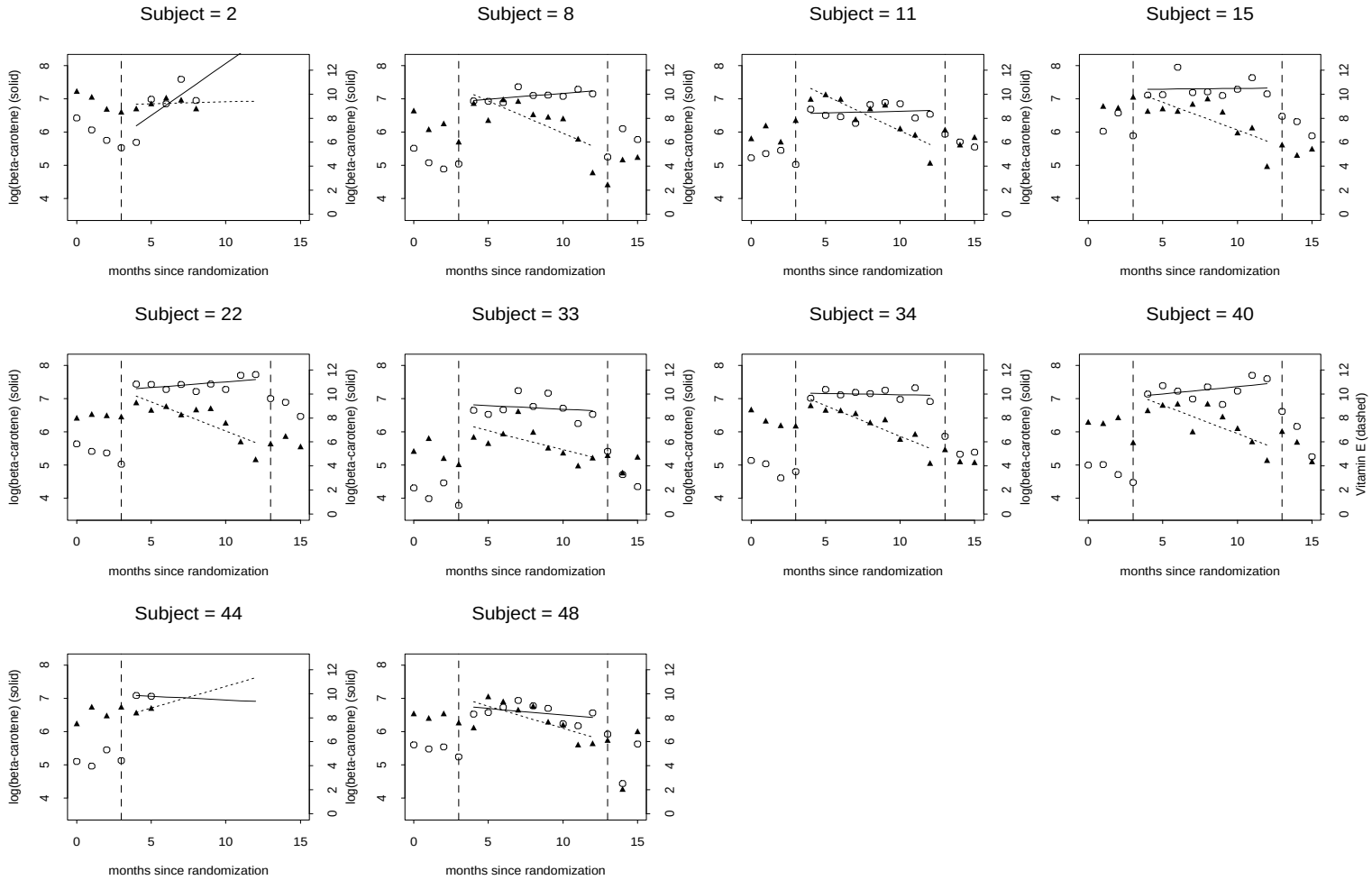
  `CHOLESTEROL` - serum cholesterol at randomization

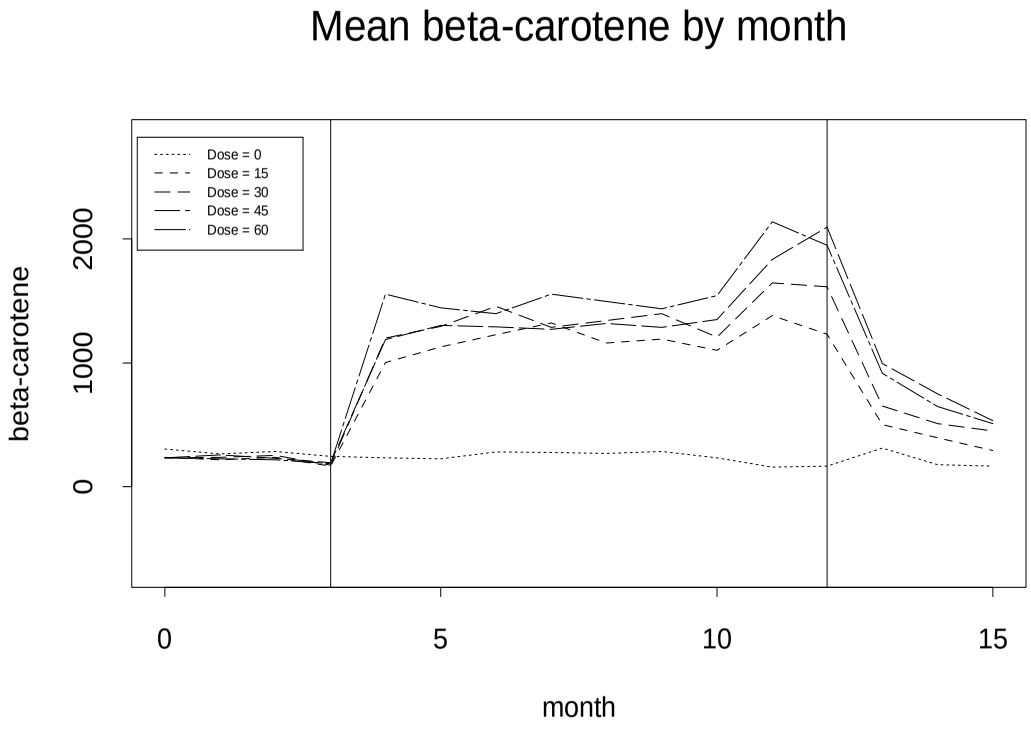
  `BODYFAT` - % bodyfat at randomization

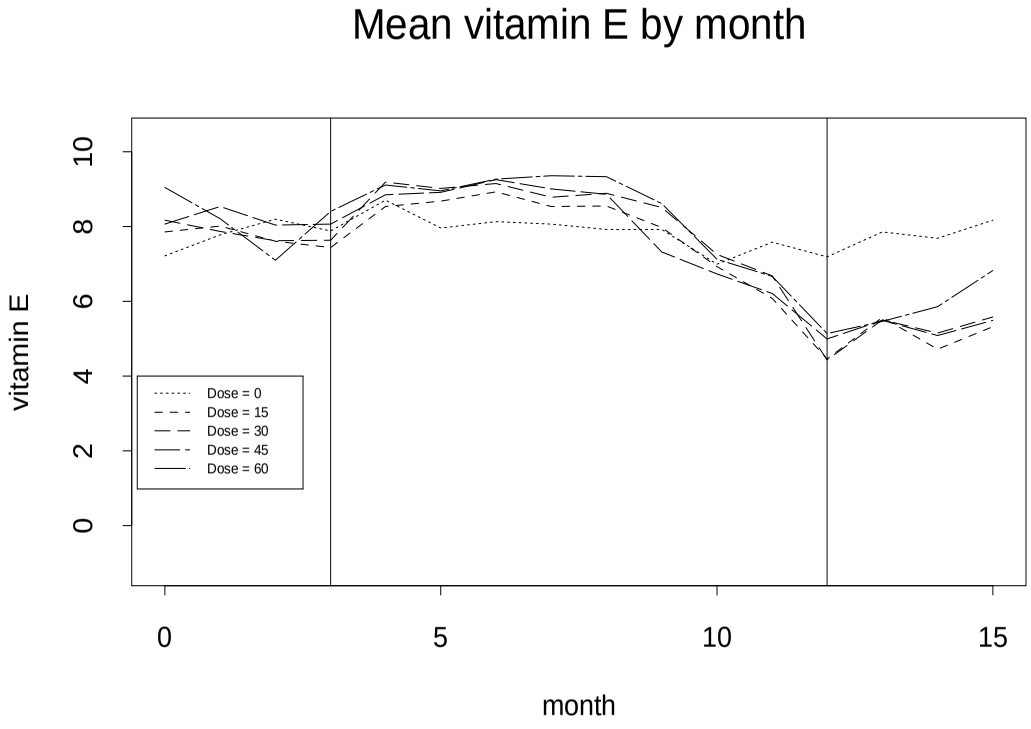
Dose = 15


Subject = 2
Subject = 8
Subject = 11
Subject = 15
Subject = 22
Subject = 33
Subject = 34
Subject = 40
Subject = 44
Subject = 48
beta-carotene (solid)
Vitamin E (dashed)
months since randomization

Dose = 15


Subject = 2
Subject = 8
Subject = 11
Subject = 15
Subject = 22
Subject = 33
Subject = 34
Subject = 40
Subject = 44
Subject = 48
log(beta-carotene) (solid)
Vitamin E (dashed)
months since randomization


Mean beta-carotene by month
beta-carotene
0
1000
2000
Dose = 0
Dose = 15
Dose = 30
Dose = 45
Dose = 60
0
5
10
15
month


Mean vitamin E by month
vitamin E
0
2
4
6
8
10
Dose = 0
Dose = 15
Dose = 30
Dose = 45
Dose = 60
0
5
10
15
month

Biostat 571 – Overview

Part III – **GLMs for Correlated Categorical Data**

approximately (4 weeks)

◦ Extend the GLM by considering a covariance structure for response vectors.

- Longitudinal data (repeated measures)
- Clustered data & "multilevel" data
- Multivariate response
- Time-series and spatial data

◦ Semi-parametric methods

- Generalized estimating equations (GEE)

- Empirical ("sandwich") variance estimates & efficiency
- Specification and estimation of covariances
- Inference

- Likelihood based methods for categorical (binary) data
  - Multivariate likelihoods for binary data
  - Generalized linear mixed models (GLMMs)
  - Prediction of random effects (empirical Bayes)
  - Clustered & Longitudinal data analysis
  - Model checking (diagnostics)
- Missing data issues!
- Additional topics?

The BIG Picture

- **Generalized linear models**
  - Models for the mean response
  - Univariate response / independent
- **Multinomial models**
  - Models for the mean response (transformed)
  - Univariate response / independent
- **Overdispersed GLMs**
  - Models for the mean response
  - Models for the variance
  - Univariate response / independent

The BIG Picture

- **General Linear Model for Correlated Data**
  - Models for the mean response (continuous)
  - Models for the covariance
  - Vector response / dependent within

- **Linear Mixed Model**
  - Models for the mean response (continuous)
  - Models for the covariance (hierarchical)
  - Vector response / dependent within

The BIG Picture

- **Marginal GLM / GEE**
  - Models for the mean response (discrete,continuous)
  - Models for the correlation
  - Vector response / dependent within

- **GLMM**
  - Models for the conditional mean response (discrete,continuous)
  - Models for the heterogeneity (hierarchical)
  - Vector response / dependent within

The BIG Picture

	SEMI-PARAMETRIC	PARAMETRIC
Overdispersion	Quasilikelihood Est. Eq. $\text{cov}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{A}^{-1}\boldsymbol{B}\ \ \boldsymbol{A}^{-1}$	beta-binomial poisson-gamma likelihood / Bayes
Continuous Resp. / linear model	WLS Est. Eq. $\text{cov}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{A}^{-1}\boldsymbol{B}\ \ \boldsymbol{A}^{-1}$	multiv. normal LMM likelihood / Bayes
Discrete Response / GLM	GEE Est. Eq. $\text{cov}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{A}^{-1}\boldsymbol{B}\ \ \boldsymbol{A}^{-1}$	multiv. dist. GLMM likelihood / Bayes

Longitudinal Data

"The basic statistical problem is that variables from a given individual are correlated over time." $(generic)$

Q: So what?

- (-) ignoring dependence can lead to invalid inference.
- (-) often limited information regarding dependence.
- (+) can observe **change** for individuals over time.
- (+) variety of statistical approaches that are available.

Longitudinal Data

"... need to account for the dependence." (*generic*)

Q: How?

1. **Choice of Model**
2. **Choice of Estimator**
3. **Choice of Summaries**

Dependent Data and Proper Variance Estimates

Let $X_{ij} = 0$ denote placebo assignment and $X_{ij} = 1$ denote active treatment.

(1) Consider (Y_{i1}, Y_{i2}) with $(X_{i1}, X_{i2}) = (0, 0)$ for $i = 1 : n$ and $(X_{i1}, X_{i2}) = (1, 1)$ for $i = (n + 1) : 2n$

$$\begin{aligned}
\hat{\mu}_0 &= \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{2} Y_{ij} \\
\hat{\mu}_1 &= \frac{1}{2n} \sum_{i=n+1}^{2n} \sum_{j=1}^{2} Y_{ij} \\
\text{var}(\hat{\mu}_1 - \hat{\mu}_0) &= \frac{1}{n}\{\sigma^2(1 + \rho)\}
\end{aligned}$$

Scenario 1

subject	control		treatment	
	time 1	time 2	time 1	time 2
ID = 101	$Y_{1,1}$	$Y_{1,2}$		
ID = 102	$Y_{2,1}$	$Y_{2,2}$		
ID = 103	$Y_{3,1}$	$Y_{3,2}$		
ID = 104			$Y_{4,1}$	$Y_{4,2}$
ID = 105			$Y_{5,1}$	$Y_{5,2}$
ID = 106			$Y_{6,1}$	$Y_{6,2}$

Dependent Data and Proper Variance Estimates

(2) Consider (Y_{i1}, Y_{i2}) with $(X_{i1}, X_{i2}) = (0, 1)$ for $i = 1 : n$ and $(X_{i1}, X_{i2}) = (1, 0)$ for $i = (n + 1) : 2n$

$$\hat{\mu}_0 = \frac{1}{2n}\left\{\sum_{i=1}^{n} Y_{i1} + \sum_{i=n+1}^{2n} Y_{i2}\right\}$$

$$\hat{\mu}_1 = \frac{1}{2n}\left\{\sum_{i=1}^{n} Y_{i2} + \sum_{i=n+1}^{2n} Y_{i1}\right\}$$

$$\mathsf{var}(\hat{\mu}_1 - \hat{\mu}_0) = \frac{1}{n}\{\sigma^2(1 - \rho)\}$$

Scenario 2

subject	control		treatment	
	time 1	time 2	time 1	time 2
ID = 101	$Y_{1,1}$			$Y_{1,2}$
ID = 102	$Y_{2,1}$			$Y_{2,2}$
ID = 103	$Y_{3,1}$			$Y_{3,2}$
ID = 104		$Y_{4,2}$	$Y_{4,1}$	
ID = 105		$Y_{5,2}$	$Y_{5,1}$	
ID = 106		$Y_{6,2}$	$Y_{6,1}$	

Dependent Data and Proper Variance Estimates

If we simply had $2n$ independent observations on treatment $(X = 1)$ and $2n$ independent observations on control then we'd obtain

$$\begin{aligned} \text{var}(\hat{\mu}_1 - \hat{\mu}_0) &= \frac{\sigma^2}{2n} + \frac{\sigma^2}{2n} \\ &= \frac{1}{n}\sigma^2 \end{aligned}$$

Q: What is the impact of dependence relative to the situation where all $(2n + 2n)$ observations are independent?

(1) $\Rightarrow$ positive dependence, $\rho > 0$, results in a loss of precision.

(2) $\Rightarrow$ positive dependence, $\rho > 0$, results in an improvement in precision!

Therefore:

- Dependent data impacts proper statements of precision.

- Dependent data may increase or decrease standard errors depending on the design.

Weighted Estimation

Consider the situation where subjects report both the number of attempts and the number of successes: (Y_i, N_i).

Examples:

live born (Y_i) in a litter (N_i)
condoms used (Y_i) in sexual encounters (N_i)
SAEs (Y_i) among total surgeries (N_i)

Q: How to combine these data from $i = 1 : m$ subjects to estimate a common rate (proportion) of successes?

Weighted Estimation

Proposal 1:

$$\hat{p}_1 = \sum_i Y_i / \sum_i N_i$$

Proposal 2:

$$\hat{p}_2 = \frac{1}{m} \sum_i Y_i / N_i$$

Simple Example:

Data : $(1, 10)$ $(2, 100)$

$$\hat{p}_1 = (2+1)/(110) = 0.030$$

$$\hat{p}_2 = \frac{1}{2}\{1/10 + 2/100\} = 0.051$$

Weighted Estimation

Note: Each of these estimators, $\hat{p}_1$, and $\hat{p}_2$, can be viewed as weighted estimators of the form:

$$\hat{p}_w = \left\{ \sum_i w_i \frac{Y_i}{N_i} \right\} / \sum_i w_i$$

We obtain $\hat{p}_1$ by letting $w_i = N_i$, corresponding to equal weight given each to binary outcome, Y_{ij}, $Y_i = \sum_{j=1}^{N_i} Y_{ij}$.

We obtain $\hat{p}_2$ by letting $w_i = 1$, corresponding to equal weight given to each subject.

Q: What's optimal?

Weighted Estimation

A: Whatever weights are closest to $1/\text{variance of } Y_i/N_i$ (stat theory called "Gauss-Markov").

- If subjects are perfectly homogeneous then

$$V(Y_i) = N_i p(1-p)$$

and $\hat{p}_1$ is best.

- If subjects are heterogeneous then, for example

$$V(Y_i) = N_i p(1-p)\{1 + (N_i - 1)\rho\}$$

and an estimator closer to $\hat{p}_2$ is best.

Summary

- Dependent data are common (and interesting!).

- Inference must account for the dependence.

- Consideration as to the choice of weighting will depend on the variance/covariance of the response variables.

Reading

◦ DHLZ Chapter 1 – examples of longitudinal studies.