

Mixed Models for Longitudinal Dichotomous Data

Don Hedeker

University of Illinois at Chicago

hedeker@uic.edu

www.uic.edu/~hedeker/long.html

Hedeker, D. (2005). Generalized linear mixed models. In B. Everitt & D. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science*. Wiley.

Chapter 9 in Hedeker, D. & Gibbons, R.D. (2006). *Longitudinal Data Analysis*. Wiley.

Logistic Regression - model that relates explanatory variables (*i.e.*, covariates) to a dichotomous dependent variable

Mixed-effects Logistic Regression - model that relates covariates to a dichotomous dependent variable, where observations are nested

- Longitudinal: repeated observations within subjects
- Clustered: subjects within clusters

models can also be recast as probit regression models

Logistic Regression Model with dichotomous x

group	x	y = response		prob	odds	logit
		0	1			
control	0	60	30	1/3	1/2	-.693
treatment	1	30	60	2/3	2	.693

$$\log \left[\frac{\text{Pr}(y_i = 1)}{1 - \text{Pr}(y_i = 1)} \right] = \beta_0 + \beta_1 x_i$$

$$\exp \beta_0 = \text{odds of response for } x = 0 \quad (30/60 = 1/2)$$

$$\hat{\beta}_0 = \log(1/2) = -.693$$

$$\exp(\beta_0 + \beta_1) = \text{odds of response for } x = 1 \quad (60/30 = 2)$$

$$\hat{\beta}_0 + \hat{\beta}_1 = \log(2) = .693$$

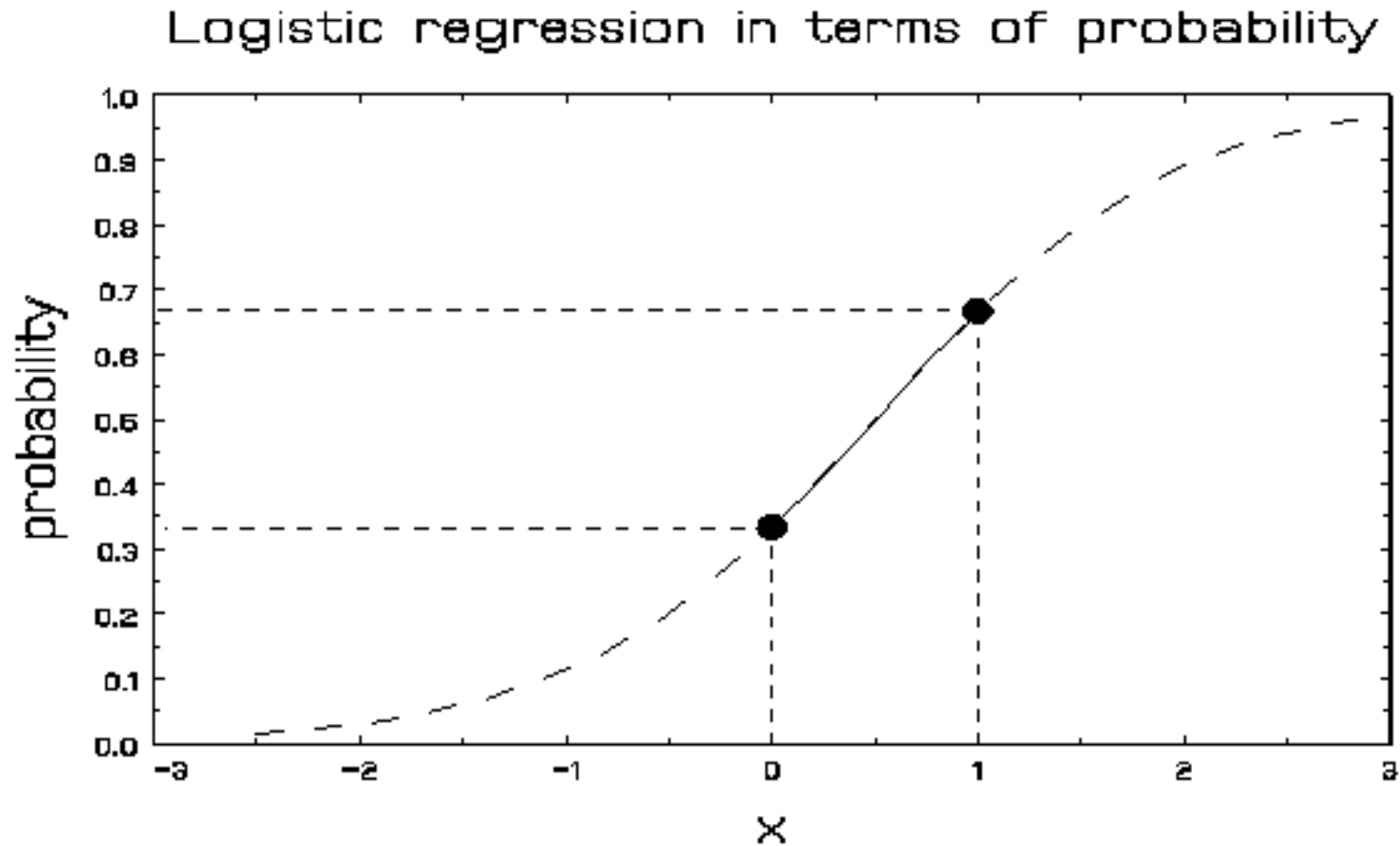
$$\hat{\beta}_1 = .693 + .693 = 1.386$$

odds ratio = ratio of odds per unit change in x

$$\begin{aligned} &= \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1)}{\exp(\hat{\beta}_0)} \\ &= \exp(\hat{\beta}_1) \\ &= \exp(1.386) = 4 \end{aligned}$$

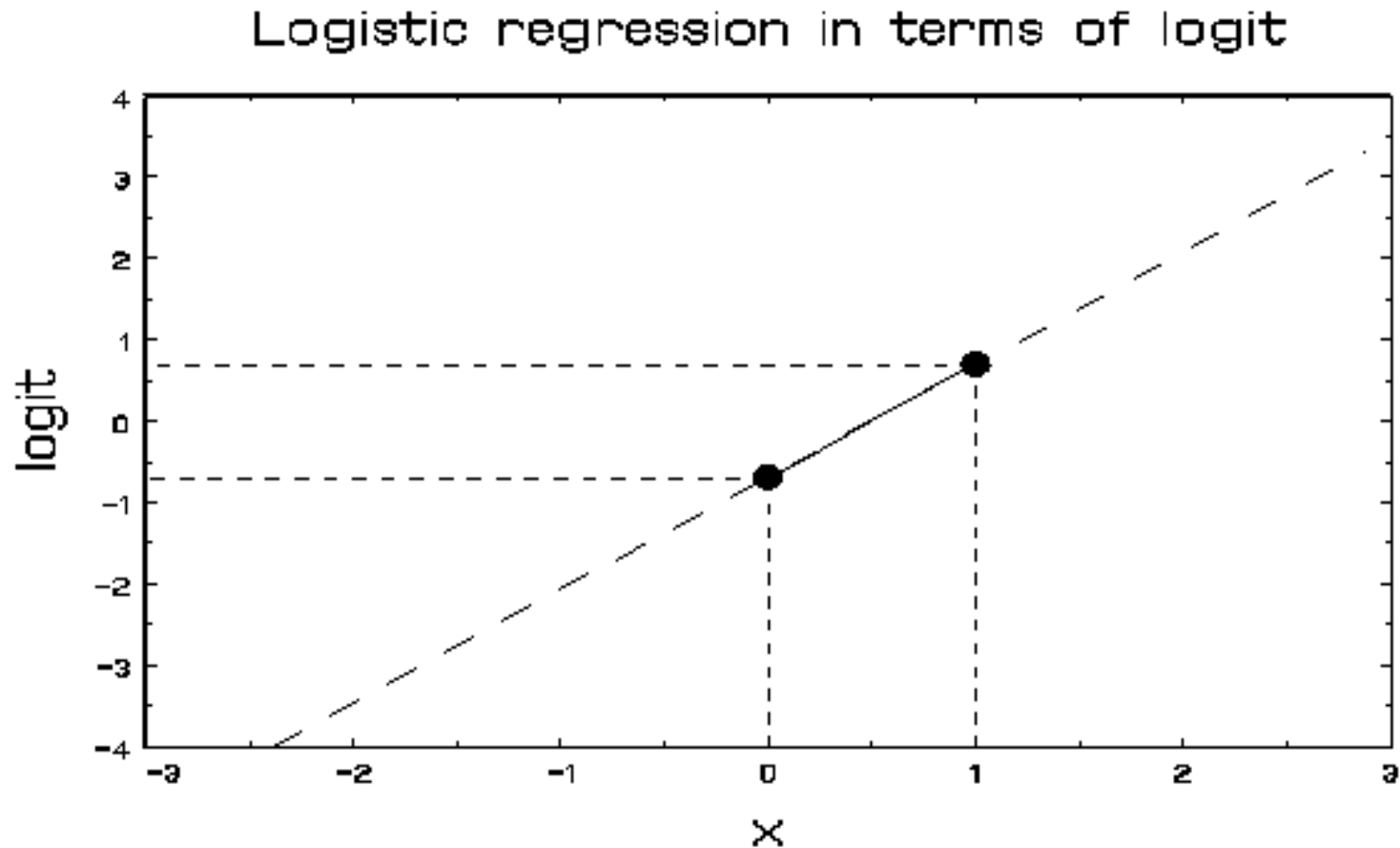
\Rightarrow Odds of response (*i.e.*, $y = 1$) are 4 times higher in treatment group, relative to control group

Model is not linear in terms of the probabilities



$$Pr(y_i = 1) = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 x_i)]}$$

Model is linear in terms of the logits



$$\log \left[\frac{\Pr(y_i = 1)}{1 - \Pr(y_i = 1)} \right] = \beta_0 + \beta_1 x_i$$

Logistic Regression Model with continuous x

age	x	y = response		prob	odds	logit
		0	1			
20-29	0	60	30	1/3	1/2	-.693
30-39	1	30	60	2/3	2	.693
40-49	2	10	80	8/9	8	2.079

$$\log \left[\frac{Pr(y_i = 1)}{1 - Pr(y_i = 1)} \right] = \beta_0 + \beta_1 x_i$$

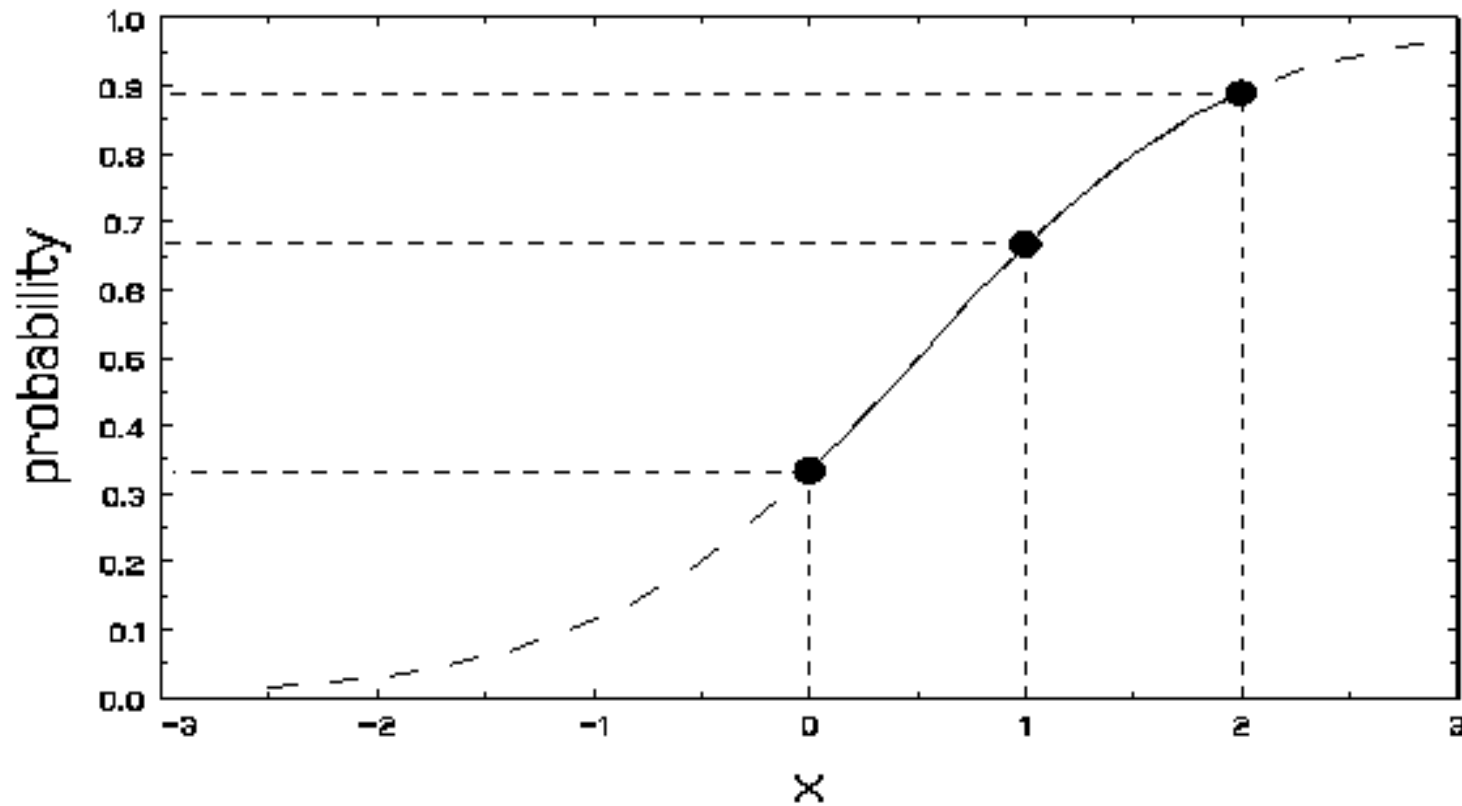
$$\hat{\beta}_0 = -.693$$

$$\begin{aligned} \hat{\beta}_1 &= \text{change in log odds w/ unit change in } x \\ &= 1.386 \end{aligned}$$

$$\begin{aligned}\text{odds ratio} &= \text{ratio of odds per unit change in } x \\ &= \exp(\hat{\beta}_1) \\ &= \exp(1.386) = 4\end{aligned}$$

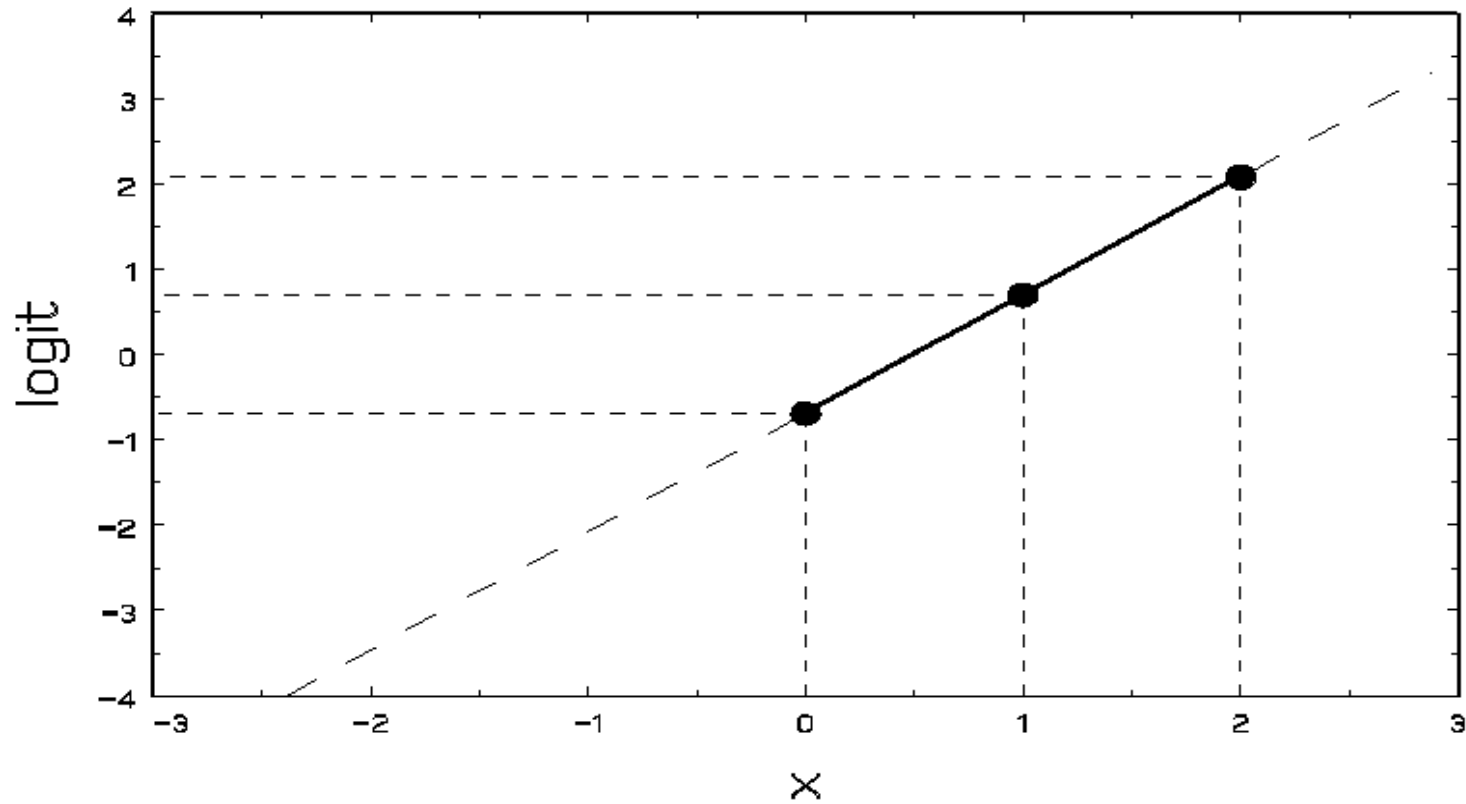
\Rightarrow Odds of response (*i.e.*, $y = 1$) are 4 times higher with each increasing decade of age

Logistic regression in terms of probability



$$Pr(y_i = 1) = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 x_i)]}$$

Logistic regression in terms of logit



$$\log \left[\frac{\text{Pr}(y_i = 1)}{1 - \text{Pr}(y_i = 1)} \right] = \beta_0 + \beta_1 x_i$$

Random-intercept Logistic Regression Model

Consider the model with p covariates for the dichotomous response y_{ij} of subject i ($i = 1, \dots, N$) at timepoint j ($j = 1, \dots, n_i$):

$$\log \left[\frac{\Pr(y_{ij} = 1)}{1 - \Pr(y_{ij} = 1)} \right] = \mathbf{x}'_{ij} \boldsymbol{\beta} + v_i$$

y_{ij} = dichotomous response of subject i at timepoint j

\mathbf{x}_{ij} = $(p + 1) \times 1$ vector of covariates

$\boldsymbol{\beta}$ = $(p + 1) \times 1$ vector of regression coefficients

v_i = random subject effects distributed $\mathcal{NID}(0, \sigma_v^2)$

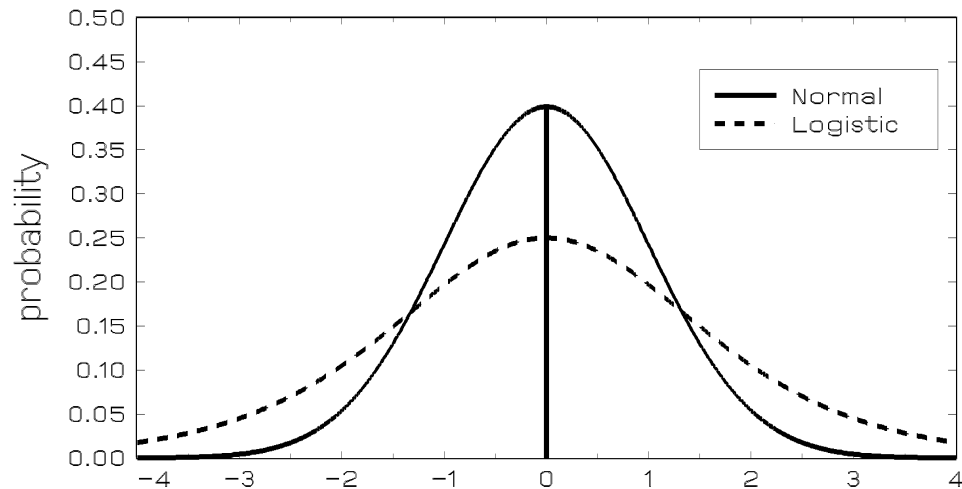
Dichotomous Response and Threshold Concept

Continuous y_{ij} - an unobservable latent variable - related to dichotomous response y_{ij} via “threshold concept”

- threshold value γ on y continuum

Response occurs $y_{ij} = 1$ if $\gamma < y_{ij}$
otherwise, a response does not occur ($y_{ij} = 0$)

Latent Distribution: Normal and Logistic pdf



The Threshold Concept in Practice

“How was your day?”

(what is your level of satisfaction today?)

- Satisfaction may be continuous, but we usually emit a dichotomous response:



Great Day!



a day ...

Model for Latent Continuous Responses

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + v_i + \varepsilon_{ij}$$

- $\varepsilon_{ij} \sim$ std normal (mean 0, variance 1): probit regression
- $\varepsilon_{ij} \sim$ std logistic (mean 0, variance $\pi^2/3$): logistic regression

Underlying latent variable

- useful way of thinking of the problem
- not an essential assumption of the model
- used for intra-class correlation

$$ICC = \frac{\sigma_v^2}{\sigma_v^2 + 1} \quad \text{for probit (equals tetrachoric if } n = 2\text{)}$$

$$= \frac{\sigma_v^2}{\sigma_v^2 + \pi^2/3} \quad \text{for logistic}$$

Scaling of regression coefficients

Fixed-effects or marginal model - β estimates from logistic are larger in absolute value than from probit by

$$\approx \sqrt{\frac{\pi^2/3}{1}} = \sqrt{\frac{\text{std logistic variance}}{\text{std normal variance}}} = 1.8$$

- Amemiya (1981) suggests 1.6, Long (1997) suggests 1.7

Random-effects model - β estimates from random-effects model are larger in abs. value than fixed-effects or marginal model by

$$\approx \sqrt{d} = \sqrt{\frac{\sigma_v^2 + \sigma^2}{\sigma^2}} = \sqrt{\frac{\text{RE variance}}{\text{FE variance}}}$$

- d = design effect in sampling literature
- Zeger et. al. (1988) $\sigma^2 = (15/16)^2\pi^2/3$ for logistic

Random-Intercept Model *Within-Subjects / Between-Subjects models*

Within-subjects model - level 1 ($j = 1, \dots, n_i$)

observed response

$$\log \left[\frac{\text{Pr}(y_{ij} = 1)}{1 - \text{Pr}(y_{ij} = 1)} \right] = b_{0i} + b_{1i} \text{Time}_{ij}$$

latent response

$$y_{ij} = b_{0i} + b_{1i} \text{Time}_{ij} + \varepsilon_{ij}$$

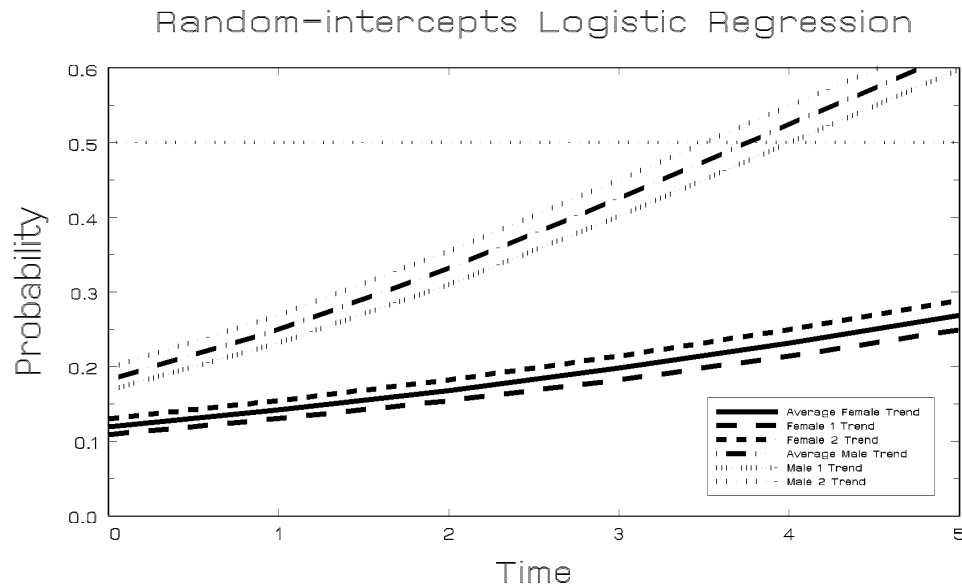
Between-subjects model - level 2 ($i = 1, \dots, N$)

$$b_{0i} = \beta_0 + \beta_2 \text{Grp}_i + v_{0i}$$

$$b_{1i} = \beta_1 + \beta_3 \text{Grp}_i$$

$$v_{0i} \sim \mathcal{NID}(0, \sigma_v^2) \quad \varepsilon_{ij} \sim \mathcal{LID}(0, \pi^2/3)$$

Random Intercept Logistic Model *in terms of probability*



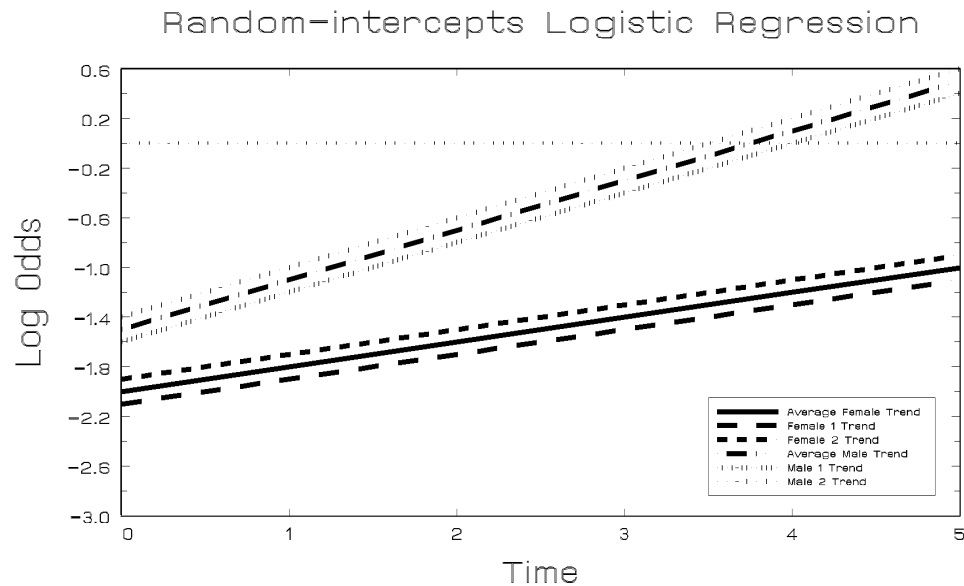
- Not linear in terms of probability

$$Pr(y_{ij} = 1) = \frac{1}{1 + \exp \left[- \left(\beta_0 + \beta_1 G_i + \beta_2 T_j + \beta_3 (G_i \times T_j) + v_{0i} \right) \right]}$$

where $G = \text{Group}$ $T = \text{Time}$

Random Intercept Logistic Model

in terms of log odds (logits)



- Linear in terms of log odds (logits)

$$\log \left[\frac{\Pr(y_{ij} = 1)}{1 - \Pr(y_{ij} = 1)} \right] = \beta_0 + \beta_1 G_i + \beta_2 T_j + \beta_3 (G_i \times T_j) + u_{0i}$$

Random Intercept and Trend Model

Within-subjects model - level 1 ($j = 1, \dots, n_i$)

latent response

$$y_{ij} = b_{0i} + b_{1i} \text{Time}_{ij} + \varepsilon_{ij}$$

Between-subjects model - level 2 ($i = 1, \dots, N$)

$$b_{0i} = \beta_0 + \beta_2 \text{Grp}_i + v_{0i}$$

$$b_{1i} = \beta_1 + \beta_3 \text{Grp}_i + v_{1i}$$

$$\begin{bmatrix} v_{0i} \\ v_{1i} \end{bmatrix} \sim \mathcal{NID} \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{v_0}^2 & \sigma_{v_0 v_1} \\ \sigma_{v_0 v_1} & \sigma_{v_1}^2 \end{bmatrix} \right\}$$

$$\varepsilon_{ij} \sim \mathcal{LID}(0, \pi^2/3)$$

Treatment-Related Change Across Time

NIMH Schizophrenia collaborative study on treatment related changes in overall severity (IMPS item # 79). Item 79, *Severity of Illness*, was scored as:

1 = normal, 2 = borderline mentally ill, 3 = mildly ill,
4 = moderately ill, 5 = markedly ill, 6 = severely ill, 7 = among the most extremely ill

The experimental design and corresponding sample sizes:

Group	Sample size at Week							<i>completers</i>
	0	1	2	3	4	5	6	
PLC (n=108)	107	105	5	87	2	2	70	65%
DRUG (n=329)	327	321	9	287	9	7	265	81%

Drug = Chlorpromazine, Fluphenazine, or Thioridazine

Main question of interest:

- Was there differential improvement for the drug groups relative to the control group?

Descriptive Statistics

Observed proportions \geq “moderately ill”

	<u>week 0</u>	<u>week 1</u>	<u>week 3</u>	<u>week 6</u>
placebo	.98	.91	.89	.71
drug	.99	.82	.66	.42

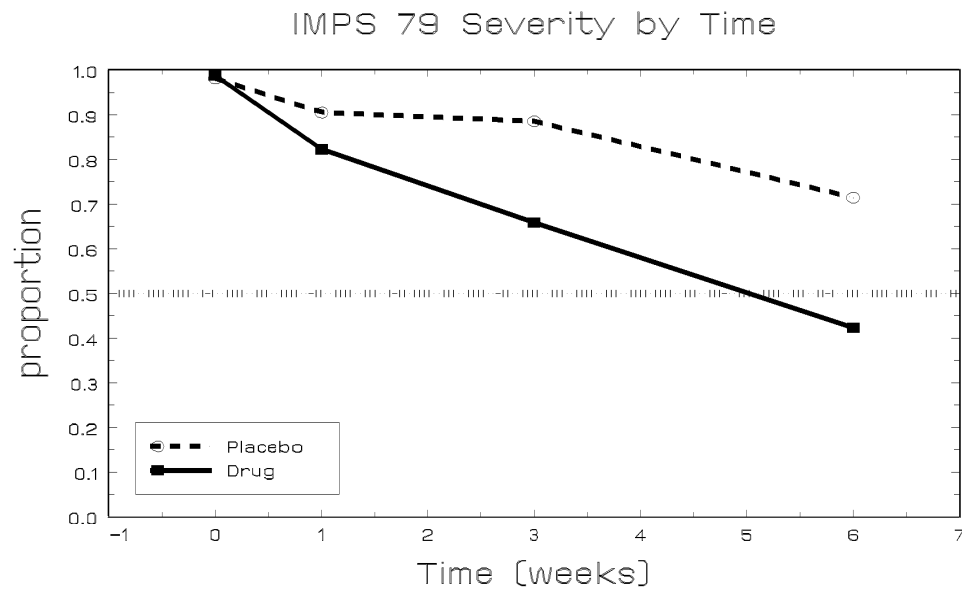
Observed odds \geq “moderately ill”

	<u>week 0</u>	<u>week 1</u>	<u>week 3</u>	<u>week 6</u>
placebo	52.5	9.50	7.70	2.50
drug	80.8	4.63	1.93	.73
<i>ratio</i>	.65	2.05	3.99	3.42

Observed log odds \geq “moderately ill”

	<u>week 0</u>	<u>week 1</u>	<u>week 3</u>	<u>week 6</u>
placebo	3.96	2.25	2.04	.92
drug	4.39	1.53	.66	-.31
<i>difference</i>	-.43	.72	1.38	1.23
exp (odds ratio)	.65	2.05	3.99	3.42

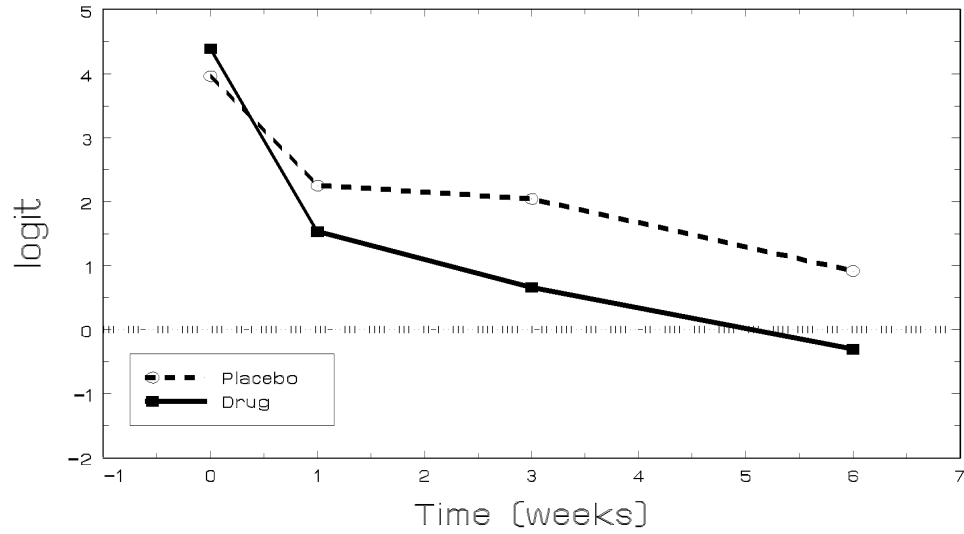
Observed Proportions across Time by Condition



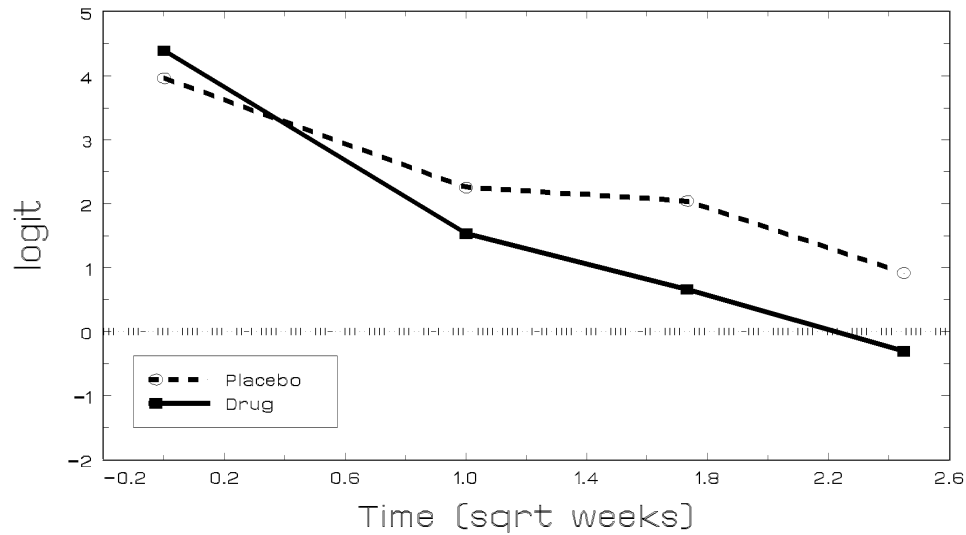
- model is not linear in terms of probabilities

Observed Logits across Time by Condition

IMPS 79 Severity by Time



IMPS 79 Severity by Time



NIMH Schizophrenia Study - Severity of Illness (N = 437)
 Logistic Regression ML Estimates - *Fixed effects model*

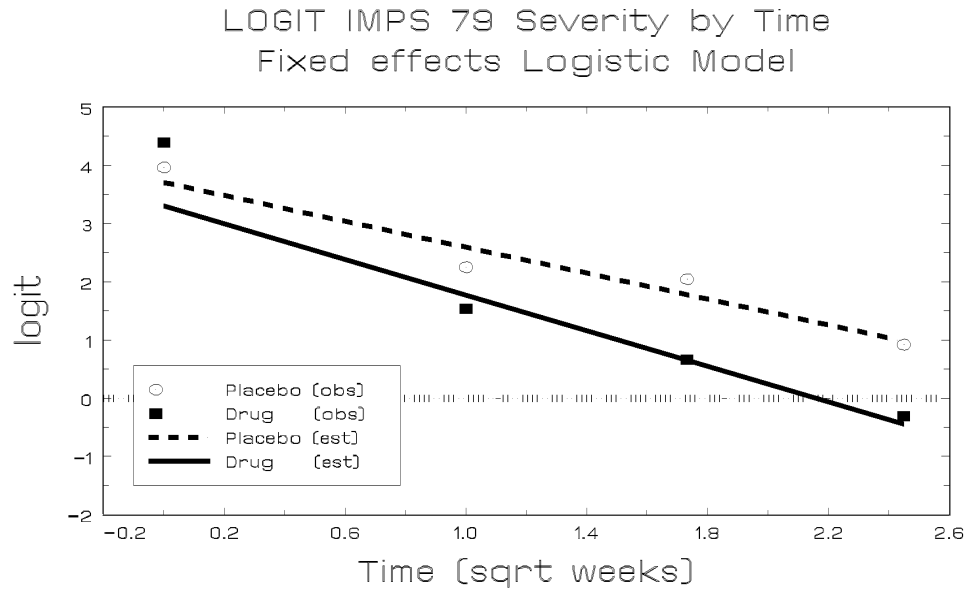
	estimates	se	z	$p <$
intercept	3.702	0.441	8.39	.001
Drug (0 = plc; 1 = drug)	-0.405	0.483	-0.84	.41
Time (sqrt week)	-1.112	0.233	-4.78	.001
Drug by Time	-0.418	0.256	-1.64	.11

$$-2 \log L = 1362.1$$

ok if data were cross-sectional longitudinal or if $\sigma_v = 0$

Fitted Logits across Time by Condition

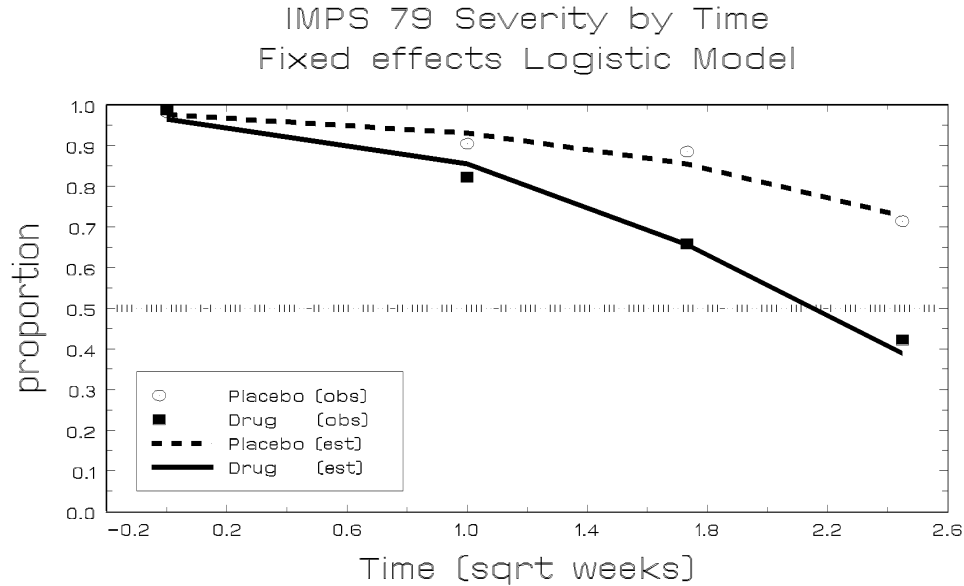
fixed-effects logistic regression model



$$\log \left[\frac{\Pr(y_{ij} = 1)}{1 - \Pr(y_{ij} = 1)} \right] = 3.70 - .41 D_i - 1.11 T_j - .42 (D_i \times T_j)$$

Fitted Proportions across Time by Condition

fixed-effects logistic regression model



$$Pr(y_{ij} = 1) = \frac{1}{1 + \exp \left[- \left(3.70 - .41 D_i - 1.11 T_j - .42 D_i T_j \right) \right]}$$

Within-Subjects / Between-Subjects components

Within-subjects model - level 1 ($j = 1, \dots, n_i$ obs)

$$\text{logit}_{ij} = b_{0i} + b_{1i}\sqrt{\text{Week}_j}$$

Between-subjects model - level 2 ($i = 1, \dots, N$ subjects)

$$b_{0i} = \beta_0 + \beta_2 \text{Grp}_i + v_{0i}$$

$$b_{1i} = \beta_1 + \beta_3 \text{Grp}_i$$

$$v_{0i} \sim \mathcal{NID}(0, \sigma_v^2)$$

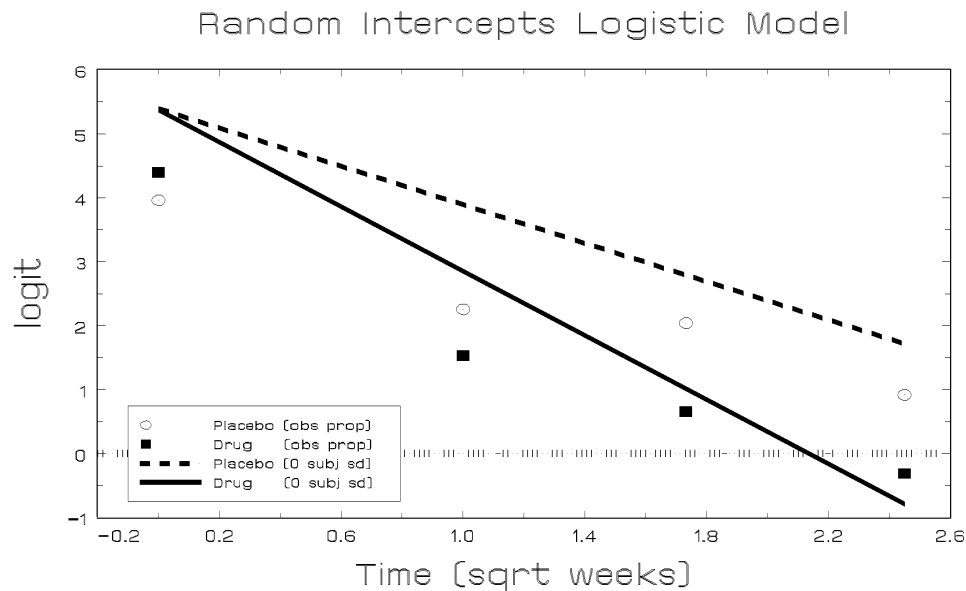
NIMH Schizophrenia Study - Severity of Illness (N = 437)
 Logistic ML Estimates (se) - *random-intercepts model*

	estimates	se	<i>z</i>	<i>p</i> <
intercept	5.365	0.626	8.57	.001
Drug (0 = plc; 1 = drug)	-0.044	0.648	-0.07	.95
Time (sqrt week)	-1.498	0.290	-5.17	.001
Drug by Time	-0.998	0.331	-3.92	.003
Intercept sd	2.082	0.214		

Intra-person correlation = $2.082^2 / (2.082^2 + \pi^2/3) = .57$

$-2 \log L = 1250.3$ $\chi_1^2 = 111.8$

Estimated (subject-specific) Logits across Time by Condition: *random-intercepts model*



$$\log \left[\frac{\text{Pr}(y_{ij} = 1)}{1 - \text{Pr}(y_{ij} = 1)} \right] = 5.37 - .04 D_i - 1.50 T_j - 1.00 (D_i \times T_j) + v_{0i}$$

$$v_{0i} \sim \mathcal{NID}(0, \hat{\sigma}_v = 2.08)$$

$\hat{\beta}$ assesses change in (conditional) logit due to \mathbf{x} for subjects with the same value of v_{0i}

Random-intercepts Logistic Regression

$$\text{logit}_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \nu_{0i}$$

- every subject has their own propensity for response (ν_{0i})
- the influence of covariates \mathbf{x} is determined controlling (or adjusting) for the subject effect
- the covariance structure, or dependency, of the repeated observations is explicitly modeled

β_0 = log odds of response for a typical subject with $\mathbf{x} = 0$ and $\nu_{0i} = 0$

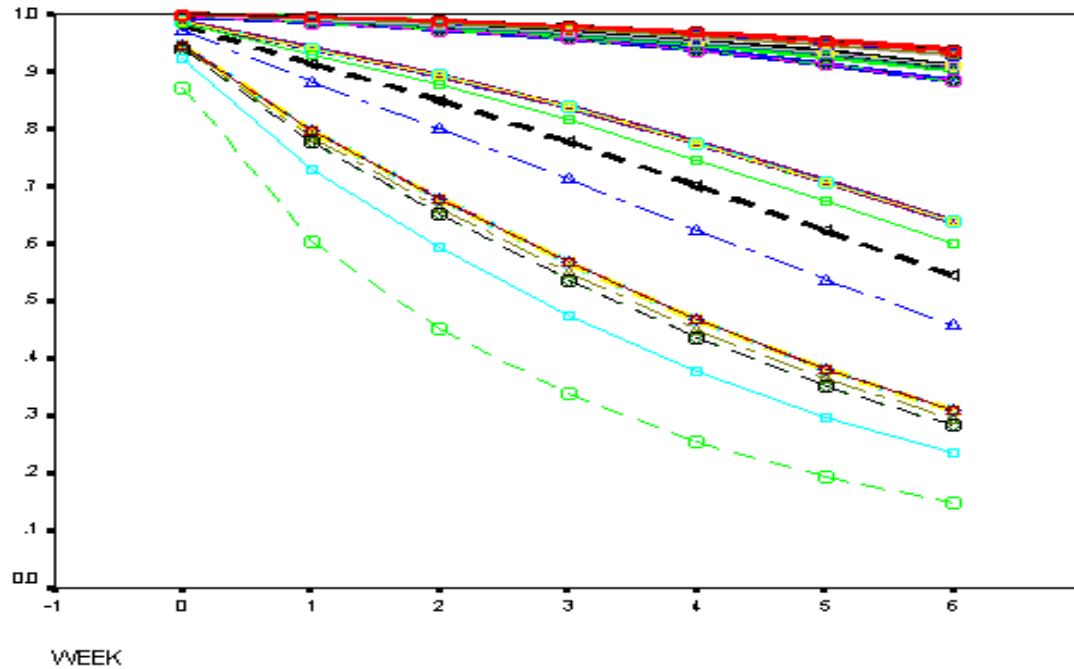
β = log odds ratio for response associated with unit changes in \mathbf{x} for the same subject value ν_{0i}
* referred to as “subject-specific”
* how a *subject's* response probability depends on \mathbf{x}

σ_ν = degree of heterogeneity across subjects in the probability of response not attributable to \mathbf{x}

- most useful when the objective is to make inference about *subjects* rather than the population average
- interest is in the heterogeneity of subjects

Estimated (subject-specific) probabilities across time

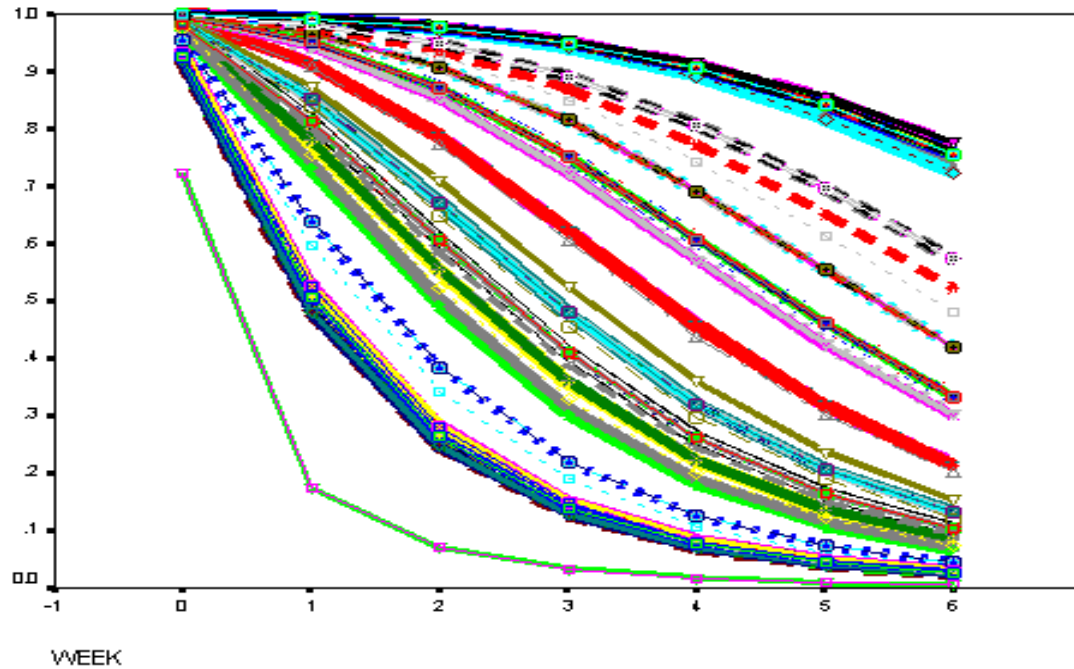
Random intercepts model - placebo group



$$P(y_{ij} = 1) = \frac{1}{1 + \exp[-(5.37 + .04 D_i - 1.50 T_j - 1.00 D_i T_j + \hat{u}_{0i})]}$$

Estimated (subject-specific) probabilities across time

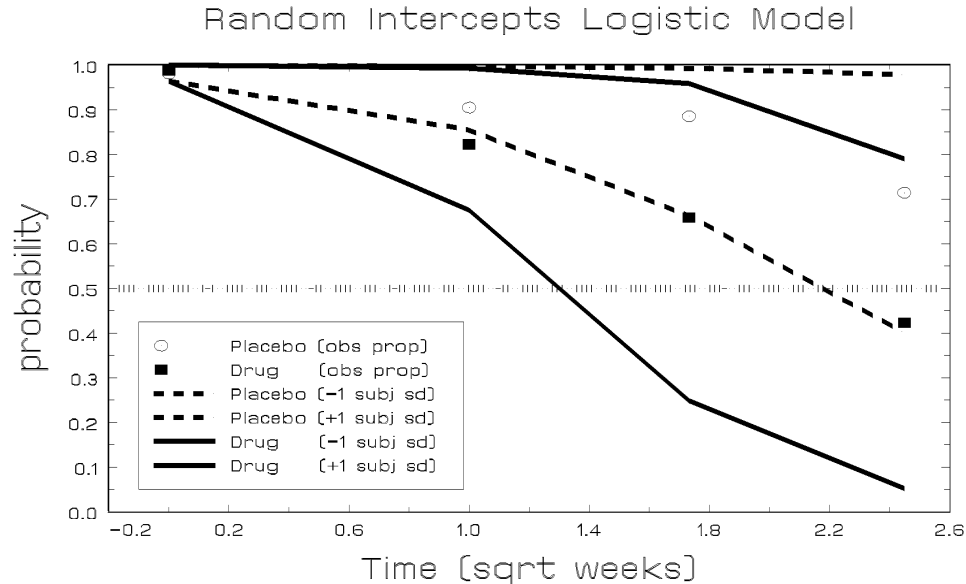
Random intercepts model - drug group



$$P(y_{ij} = 1) = \frac{1}{1 + \exp[-(5.37 - .04 D_i - 1.50 T_j - 1.00 D_i T_j + \hat{v}_{0i})]}$$

Estimated Subject-Specific Probabilities

random-intercepts logistic regression model



$$Pr(y_{ij} = 1) = \frac{1}{1 + \exp \left[- \left(5.37 - .04 D_i - 1.50 T_j - 1.00 D_i T_j + v_{0i} \right) \right]}$$

$$\text{where } v_{0i} = \begin{cases} -1\sigma_v \\ 1\sigma_v \end{cases} \text{ and } \hat{\sigma}_v = 2.08$$

Model fit of observed marginal proportions

1. $\hat{\mathbf{y}}_i = \mathbf{X}_i \hat{\boldsymbol{\beta}}$

2. calculate marginalization factor

$$\hat{s} = \sqrt{\hat{d}} = \sqrt{(\hat{\sigma}_v^2 + \sigma^2)/\sigma^2} = \sqrt{\hat{\sigma}_v^2/\sigma^2 + 1}$$

- $\sigma = 1$ for probit or $\sigma = \pi/\sqrt{3}$ for logistic
- \hat{d} is the design effect in the sampling literature

3. marginalize $\hat{\mathbf{z}}_i = \hat{\mathbf{y}}_i / \hat{s}$

4. $\hat{\mathbf{p}}_i = \Phi(\hat{\mathbf{z}}_i)$ for probit and $\hat{\mathbf{p}}_i = \Psi(\hat{\mathbf{z}}_i)$ for logistic, Φ represents the normal cdf and Ψ the logistic cdf, *i.e.*, $1/[1 + \exp(-\hat{\mathbf{z}}_i)]$

notes:

- In practice, for logistic, $(15\pi)/(16\sqrt{3})$ works better than $\pi/\sqrt{3}$ as σ (Zeger *et al.*, 1988, Biometrics)
- Logistic is approximate; relies on cumulative Gaussian approximation to the logistic function
- For multiple random effects, calculate marginalization vector

$$\hat{\mathbf{s}} = \frac{1}{\sigma} \left[\text{Diag}(\hat{V}(\mathbf{y}_i)) \right]^{1/2}$$

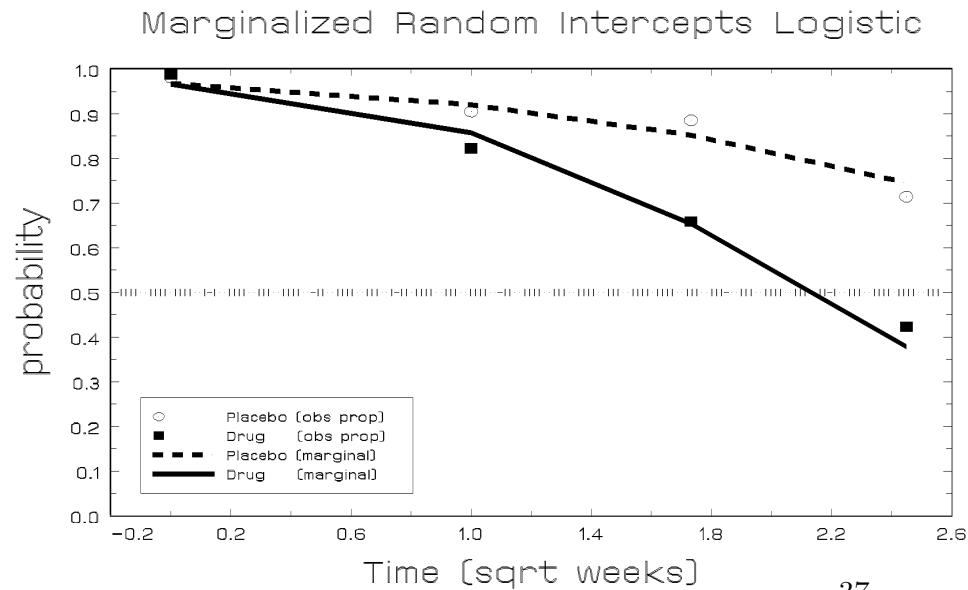
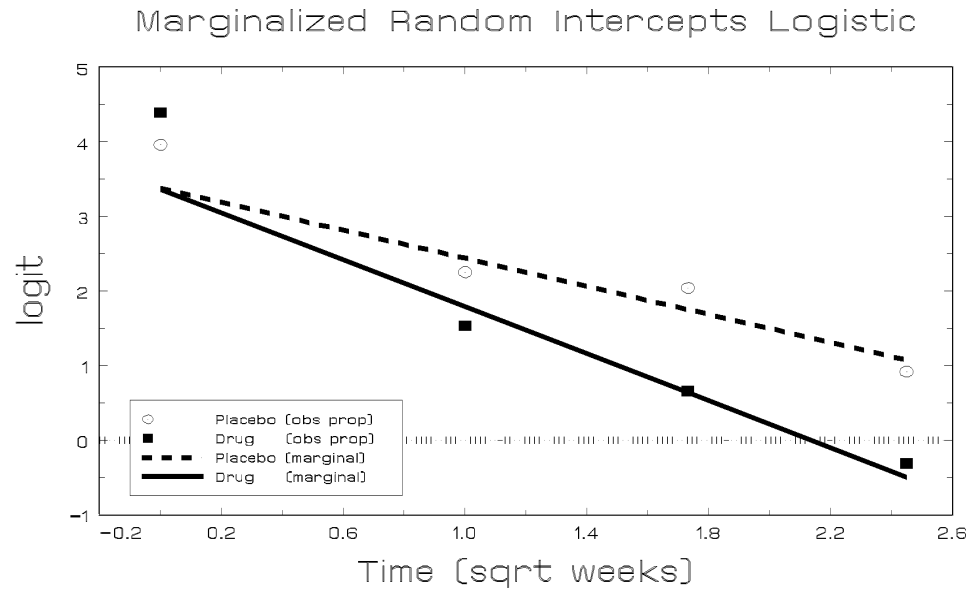
$$- \hat{V}(\mathbf{y}_i) = \mathbf{Z}_i \hat{\Sigma}_v \mathbf{Z}_i' + \sigma^2 \mathbf{I}_i$$

- \mathbf{Z}_i = design matrix for random effects

and perform element-wise division

$$\hat{\mathbf{z}}_i = \hat{\mathbf{y}}_i / \cdot \hat{\mathbf{s}}$$

Estimated Marginal Logits and Probabilities



SAS NLMIXED code: SCHZBINL.SAS

```
DATA one; INFILE 'c:\mixdemo\schizx1.dat';
INPUT id imps79 imps79b imps79o int tx week sweek txswk ;

/* get rid of observations with missing values */
IF imps79 > -9;

PROC FORMAT;
VALUE imps79b 0 = 'le mild' 1 = 'ge moderate';
VALUE tx 0 = 'placebo' 1 = 'drug';

/* fixed-effects logistic regression model */
PROC LOGISTIC DESCENDING;
MODEL imps79b = tx sweek tx*sweek;

/* random intercept logistic regression model */
PROC NLMIXED ;
PARMS b0=3.70 b1=-.40 b2=-1.11 b3=-.42 sd=1;
z = b0 + b1*tx + b2*sweek + b3*tx*sweek + u;
IF (imps79b=1) THEN
p = 1 / (1 + EXP(-z));
ELSE
p = 1 - (1 / (1 + EXP(-z)));
l1 = LOG(p);
MODEL imps79b ~ GENERAL(l1);
RANDOM u ~ NORMAL(0,sd*sd) SUBJECT=id OUT=ebest1;
ESTIMATE 'icc' sd*sd/(3.289868134+sd*sd);
```

SAS IML code: SCHZBFIT1.SAS

```
TITLE1 'nimh schizophrenia data - estimated marginal probabilities';
PROC IML;
/* Results from nlmixed analysis: random intercept model */;

/* covariate matrices for placebo and drug groups */;
x0 = { 1 0 0.00000 0,
       1 0 1.00000 0,
       1 0 1.73205 0,
       1 0 2.44949 0};
x1 = { 1 1 0.00000 0.00000,
       1 1 1.00000 1.00000,
       1 1 1.73205 1.73205,
       1 1 2.44949 2.44949};

/* nlmixed estimates of covariate effects and random effect standard deviation */;
beta = {5.365, -0.044, -1.498, -0.998};
sd    = {2.082};

/* marginalization of person-specific estimates */;
pi    = 3.141592654;
nt    = 4;
ivec  = J(nt,1,1);
zvec  = J(nt,1,1);
evec  = (15/16)**2 * (pi**2)/3 * ivec;
```

```
/* nt by nt matrix with evec on the diagonal and zeros elsewhere */;
emat = DIAG(evec);

/* variance-covariance matrix of underlying latent variable */;
vary = zvec * sd * T(sd) * T(zvec) + emat;

/* marginalization factor */;
sdy = SQRT(VECDIAG(vary) / VECDIAG(emat));

z0 = (x0*beta) / sdy ;
z1 = (x1*beta) / sdy;

grp0 = 1 / ( 1 + EXP(0 - z0));
grp1 = 1 / ( 1 + EXP(0 - z1));

print 'random intercept model';
print 'marginalization of person-specific estimates';
print 'marginal prob for group 0 - response' grp0 [FORMAT=8.4];
print 'marginal prob for group 1 - response' grp1 [FORMAT=8.4];
```

Random intercept and trend model

within-subjects / between-subjects components

within-subjects model - level 1 ($j = 1, \dots, n_i$ obs)

$$\text{logit}_{ij} = b_{0i} + b_{1i}\sqrt{\text{Week}_j}$$

between-subjects model - level 2 ($i = 1, \dots, N$ subjects)

$$b_{0i} = \beta_0 + \beta_2 \text{Grp}_i + v_{0i}$$

$$b_{1i} = \beta_1 + \beta_3 \text{Grp}_i + v_{1i}$$

$$\mathbf{v}_i \sim \mathcal{NID}(\mathbf{0}, \Sigma_v)$$

Logistic ML Estimates (se) - random intercept and trend model

	estimates	se	z	$p <$
intercept	5.814	0.909	6.40	.001
Drug (0 = plc; 1 = drug)	0.274	0.735	0.37	.71
Time (sqrt week)	-1.390	0.460	-3.02	.003
Drug by Time	-1.556	0.464	-3.35	.001

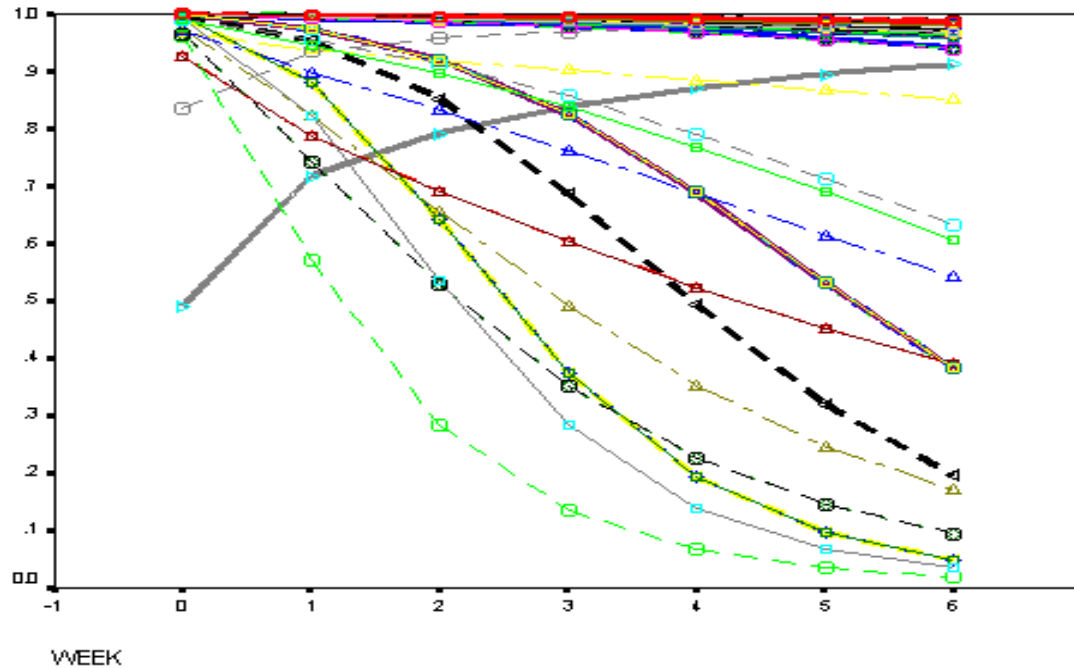
Variance-covariance terms

Intercept var	6.504	2.726		
Int-Time covar	-1.802	1.106	$(r_{v_0v_1} = -.43)$	
Time var	2.682	0.985		

$-2 \log L = 1228.4, \chi_2^2 = 21.9, p < .001$

Estimated (subject-specific) probabilities across time

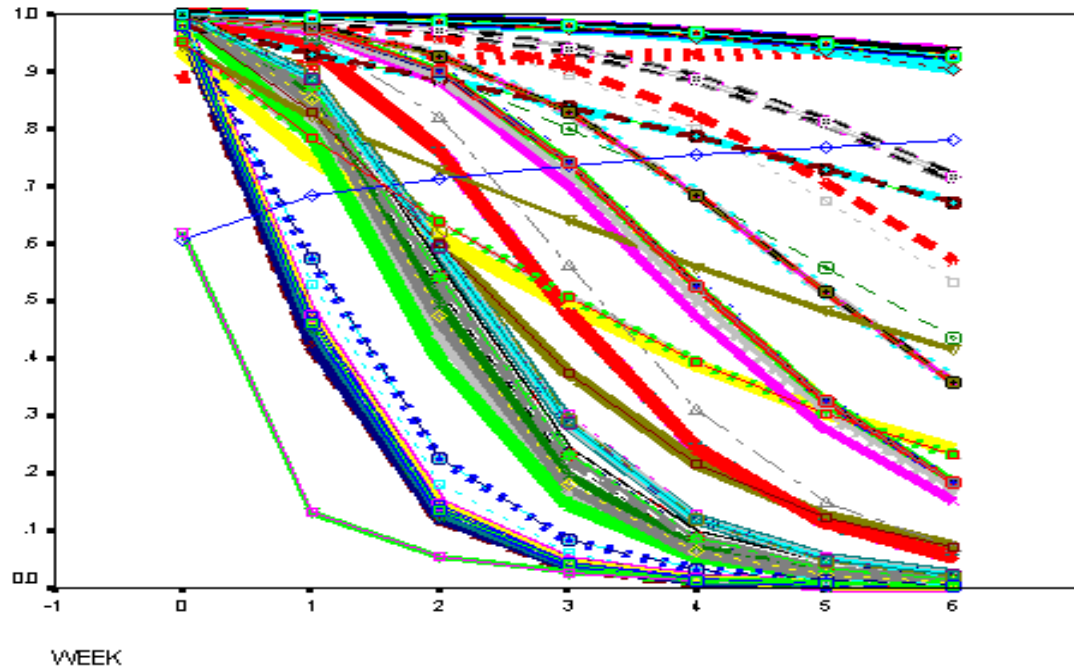
Random intercepts and trends model - placebo group



$$P(y_{ij} = 1) = \frac{1}{1 + \exp[-(5.83 + .23 D_i - 1.44 T_j - 1.49 D_i T_j + \hat{v}_{0i} + \hat{v}_{1i} T_j)]}$$

Estimated (subject-specific) probabilities across time

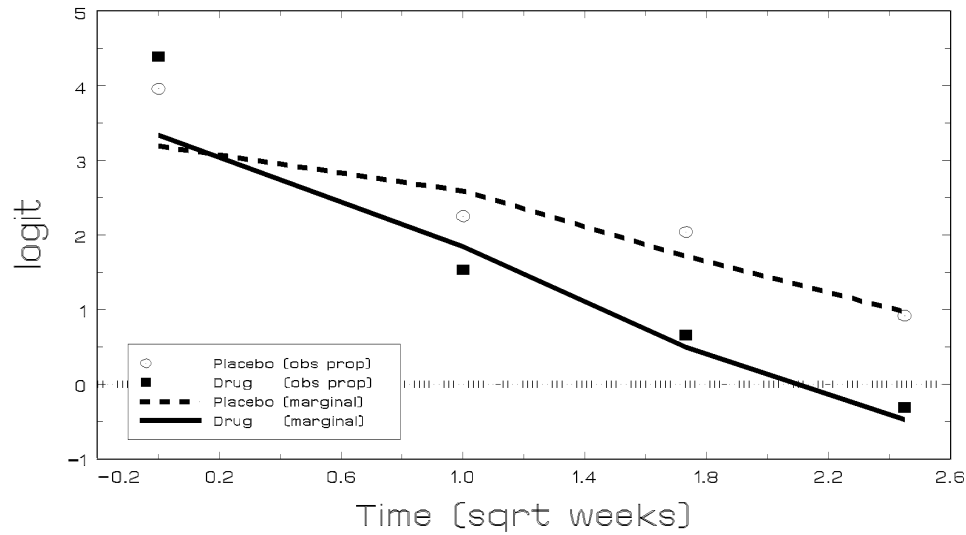
Random intercepts and trends model - drug group



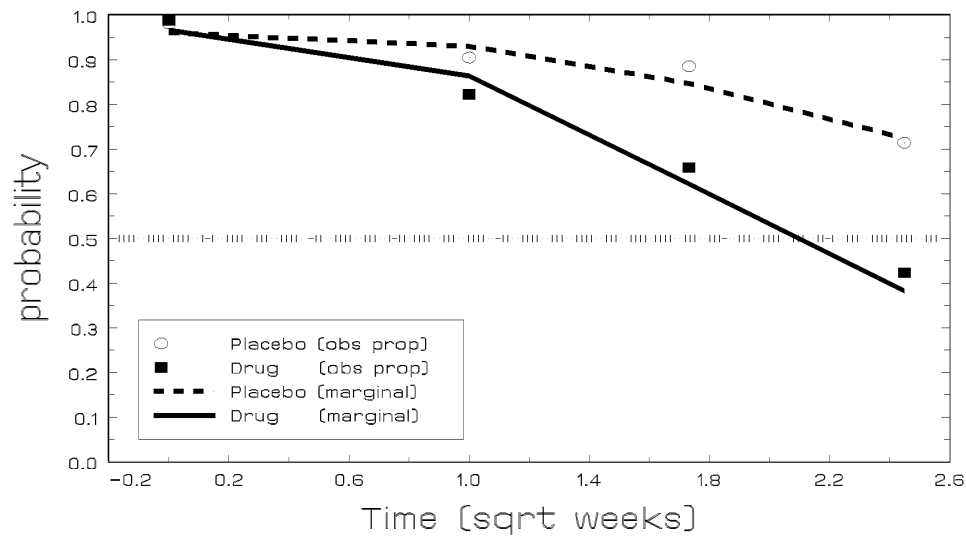
$$P(y_{ij} = 1) = \frac{1}{1 + \exp[-(5.83 + .23 D_i - 1.44 T_j - 1.49 D_i T_j + \hat{v}_{0i} + \hat{v}_{1i} T_j)]}$$

Estimated Marginal Logits and Probabilities

Marginalized Random Int & Trend Logistic



Marginalized Random Int & Trend Logistic



SAS NLMIXED code: random-trend logistic regression
included in SCHZBINL.SAS syntax file

```
/* logistic random-trend model */
PROC NLMIXED ;
PARMS b0=5.365 b1=-0.044 b2=-1.498 b3=-0.998 v0=4.335 c01=0 v1=1;
z = b0 + b1*tx + b2*sweek + b3*tx*sweek + u0 + u1*sweek;
IF (imps79b=1) THEN
p = 1 / (1 + EXP(-z));
ELSE
p = 1 - (1 / (1 + EXP(-z)));
l1 = LOG(p);
MODEL imps79b ~ GENERAL(l1);
RANDOM u0 u1 ~ NORMAL([0,0], [v0,c01,v1]) SUBJECT=id OUT=ebest2;
ESTIMATE 're corr' c01/SQRT(v0*v1);
RUN;
```

SAS IML code: SCHZBFIT2.SAS

```
TITLE1 'nimh schizophrenia Data - estimated marginal probabilities';
PROC IML;
/* results from nlmixed analysis: random intercept & trend model */;

/* covariate matrices for placebo and drug groups */;
x0 = { 1 0 0.00000 0,
      1 0 1.00000 0,
      1 0 1.73205 0,
      1 0 2.44949 0};
x1 = { 1 1 0.00000 0.00000,
      1 1 1.00000 1.00000,
      1 1 1.73205 1.73205,
      1 1 2.44949 2.44949};

/* nlmixed estimates of covariate effects and random effect variance-covariance matrix */;
beta = { 5.814, 0.274, -1.390, -1.556};
varu = {6.504 -1.802,
      -1.802 2.682};

/* marginalization of person-specific estimates */;
pi = 3.141592654;
nt = 4;
ivec = J(nt,1,1);
zmat = {1 0.00000,
      1 1.00000,
      1 1.73205,
      1 2.44949};
```

```

evec = (15/16)**2 * (pi**2)/3 * ivec;

/* nt by nt matrix with evec on the diagonal and zeros elsewhere */;
emat = DIAG(evec);

/* variance-covariance matrix of underlying latent variable */;
vary = zmat * varu * T(zmat) + emat;

/* marginalization factor */;
sdy = SQRT(VECDIAG(vary) / VECDIAG(emat));

z0 = (x0*beta) / sdy ;
z1 = (x1*beta) / sdy;

grp0 = 1 / ( 1 + EXP(0 - z0));
grp1 = 1 / ( 1 + EXP(0 - z1));

print 'random intercept and trend model';
print 'marginalization of person-specific estimates';
print 'marginal response probability for group 0' grp0 [FORMAT=8.4];
print 'marginal response probability for group 1' grp1 [FORMAT=8.4];

```

Logistic GEE as marginal model

$$\text{logit}_{ij} = \mathbf{x}'_{ij} \boldsymbol{\beta}$$

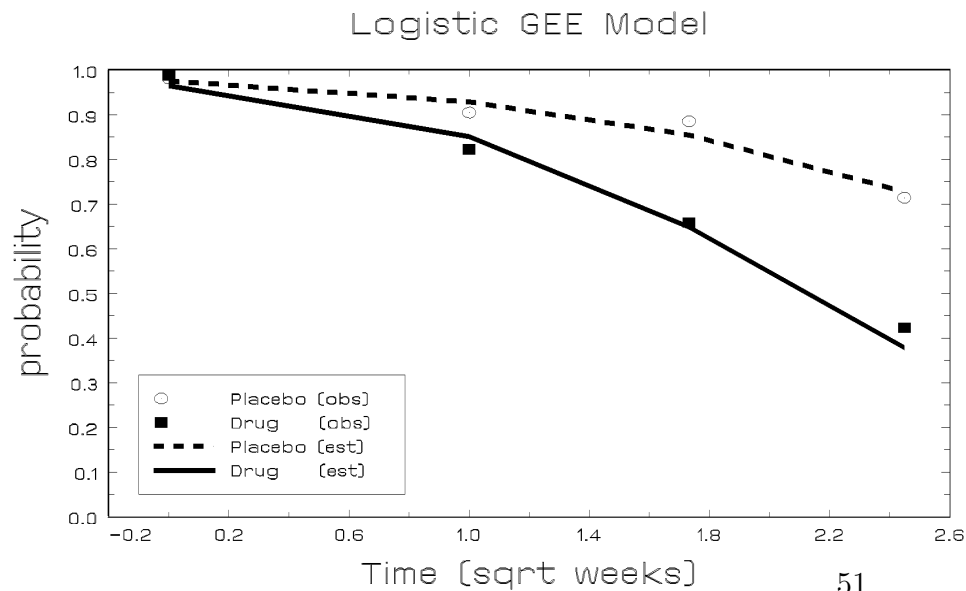
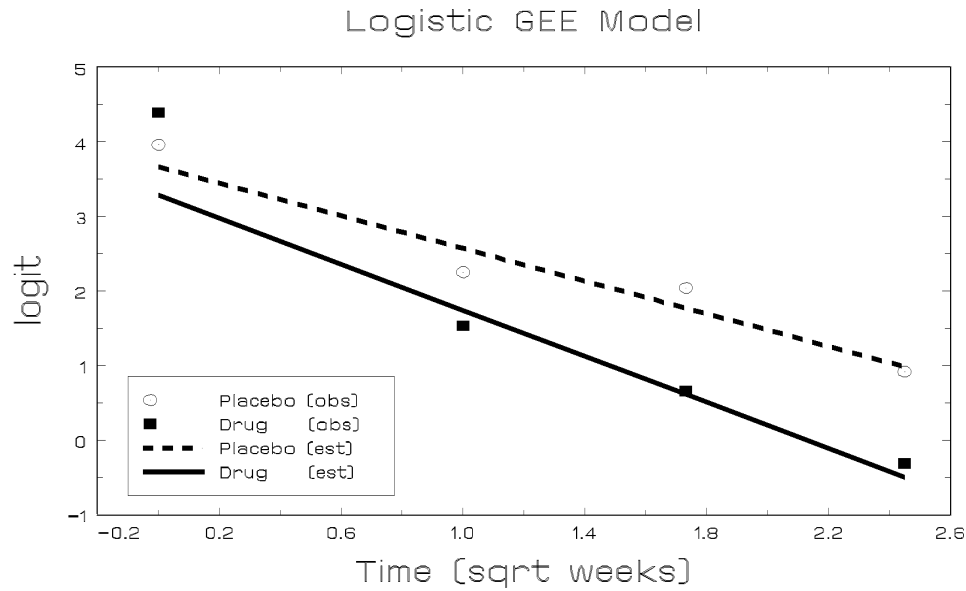
- Working correlation of repeated observations exchangeable (all are equal), AR(1), banded (m -dependent), unstructured
- robust standard errors
- does not include any subject-specific (random) effects, does not focus on heterogeneity
 - $\beta_0 = \log$ odds of response among sub-population with $\mathbf{x} = 0$
 - $\boldsymbol{\beta} = \log$ odds ratio for response associated with unit changes in \mathbf{x} in the population of subjects
- $\exp(\boldsymbol{\beta}) = \text{ratio of population frequencies}$
 - referred to as “population-averaged”

NIMH Schizophrenia Study - Severity of Illness (N = 437)
 Logistic Regression GEE - exchangeable correlation structure

	GEE estimates	se	z	$p <$
intercept	3.661	0.485	7.54	.001
Drug (0 = plc; 1 = drug)	-0.381	0.521	-0.73	.46
Time (sqrt week)	-1.094	0.252	-4.35	.001
Drug by Time	-0.449	0.269	-1.67	.10

- non-significant drug by time interaction
- working corr based on data from 7 timepts (weeks 0 to 6)
- several have little data (wks 2, 4, 5) & wk 0 is near-constant
- very poorly estimated working correlation matrix
- analysis of 4 primary timepts and UN working corr yields significant interaction ($p < .047$)

Estimated Marginal Logits and Probabilities



SAS GENMOD code: GEE logistic regression - SCHZGEE.SAS

```
DATA one; INFILE 'c:\mixdemo\schizx1.dat';
INPUT id imps79 imps79b imps79o int tx week sweek txswk;

/* get rid of observations with missing values */
IF imps79 > -9;

/* get rid of weeks with very few observations */
IF week EQ 0 or week EQ 1 OR week EQ 3 OR week EQ 6;

PROC FORMAT;
VALUE imps79b 0 = 'le mild' 1 = 'ge moderate';
VALUE tx 0 = 'placebo' 1 = 'drug';

/* gee logistic regression model: unstructured */
PROC GENMOD DESCENDING;
CLASS id week;
MODEL imps79b = tx sweek txswk / LINK=LOGIT DIST=BIN;
REPEATED SUBJECT=id / WITHIN=week CORRW TYPE=UN;
RUN;
```

Conclusions - mixed-effects logistic regression models useful for incomplete longitudinal dichotomous data

- can handle subjects measured incompletely or at different timepoints (missing data assumed MAR)
- degree of within-subjects variation on dichotomous outcome is important to consider (might have 3-timepoint study where 90% of subjects have same response across timepoints)
- subject-specific (or conditional) interpretation of regression coefficients
- generalizations to other categorical outcomes
 - ordinal outcomes - mixed-effects ordinal logistic regression
 - * proportional odds model
 - * partial or non-proportional odds model
 - nominal outcomes - mixed-effects nominal logistic regression