Faculty of Health Sciences

Models for binary data

Analysis of repeated measurements 2015

Julie Lyng Forman & Lene Theil Skovgaard

Department of Biostatistics, University of Copenhagen

Program for day 6

Models for binary repeated measurements

Population average models (PA)

- Estimation of **population** risk (depending on covariates).
- E.g. disease prevalence in the population.
- Using generalized estimating equation to account for replicates/clustering.

Subject specific models (SS)

- Estimation of **individual** risk (depending on covariates).
- E.g. chages over time, effect of treatment.
- Using generalized linear mixed models.

Suggested reading: Fitzmaurice et al. (2011): chapters 11–16.

Case studies

Longitudinal data

- Amenorrhea in contracepting women
- Randomized clinical study
- Balanced design with two groups and 4 times of follow-up

Clustered data (variance components)

- Maybe next time . . .

Challenges

- Visualization: spaghettiplots are no good :(
- Differences bewteen the PA and the SS model.
- Interpretation of random effects in the SS models.
- Missing data
- Too small (or too large) data sets.

Outline

Repetition: Binary data and logistic regression

Describing dependence between two binary outcomes

Population average models

Missing data in PA models

Subject specific models (GLMMs)

Comparison of PA and SS models

Examples of binary outcomes

Disease, symptom, or side effect.

- E.g. infection, headache, or amenorrhea.
- Recorded as **yes:** $Y = 1$, **no:** $Y = 0$

in a prespecified time interval.

Dead or alive . . . ?

- Not relevant in longitudinal data, you only die once! Better use **survival analysis** in this case.
- May be sensible with clustered data, if all subjects have similar follow-up times.

The Bernoulli distribution

A binary variable Y has a Bernoulli distribution, meaning that

$$P(Y = 1) = p, \quad \text{and} \quad P(Y = 0) = 1 - p$$

- This probability parameter p fully specifies the distribution.
- The mean is $E(Y) = \mu(p) = p$.
- The variance is $\mathrm{Var}(Y) = \sigma^2(p) = p(1 - p)$.

  Note: We don't have to model the variance; it is determined by the mean.

How do we model e.g. risk of disease as depending on covariates?

Case study: Amenorrhea

1151 women were randomized in two groups, receiving a contracepting drug in either

- low dose of 100 mg (`trt=0`)
- high dose of 150 mg (`trt=1`)

All women received injections every 90 days, four times in total.

Following each injection period it was recorded whether the woman had experienced *amenorrhea* (a suspected side effect of the drug). Note that no recording was made at baseline.

Objective: *How common is amenorrhea as a side effects of the contraception? Is the risk of amenorrhea higher with high dose than with low?*

Amenorrhea: relative frequencies

```
                                 Analysis Variable : amenorrhea

                                      N                   N
        dose             time       Obs         N      Miss             Mean         Variance
-----------------------------------------------------------------------------------------------
           0                1       576       576         0        0.1857639        0.1515187

                            2       576       477        99        0.2620545        0.1937882

                            3       576       409       167        0.3887531        0.2382065

                            4       576       361       215        0.5013850        0.2506925

           1                1       575       575         0        0.2052174        0.1633874

                            2       575       476        99        0.3361345        0.2236179

                            3       575       389       186        0.4935733        0.2506029

                            4       575       353       222        0.5354108        0.2494527
-----------------------------------------------------------------------------------------------
```

Note: missing data

Logistic regression

For outcomes that are binary a linear model is not reasonable since the mean is constrained to the $[0;1]$-interval.

Instead we use logistic regression to model the log-odds of an outcome as dependent on covariates

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

The probability $p = P(Y=1) = E(Y)$ can be recovered using back-transformation

$$E(Y) = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)}$$

Odds ratios

Regression parameters in logistic regression must be transformed to odds ratios for interpretation.

- The odds ratio: OR $= \exp(\beta)$ describes the increase in odds when X is increased by 1 unit.
- **Example:** OR $= 1.13$ higher odds of amenorrhea for high dose at first occation.
- OR approximates the relative risk when the probability of the outcome is small
- When the probabilty is moderate or high the interpretation of the OR is more difficult, a large OR may reflect at difference in probabilities that is small in absolute terms.

Outline

Repetition: Binary data and logistic regression

Describing dependence between two binary outcomes

Population average models

Missing data in PA models

Subject specific models (GLMMs)

Comparison of PA and SS models

Cross tables

E.g. Same subject at two different occations, e.g. 1 and 2.

low dose group

```
amenorrhea1     amenorrhea2

Frequency|
Row Pct  |       0|       1|  Total
---------+--------+--------+
       0 |    322 |     71 |    393
         |  81.93 |  18.07 |
---------+--------+--------+
       1 |     30 |     54 |     84
         |  35.71 |  64.29 |
---------+--------+--------+
Total         352      125      477
```

high dose group

```
amenorrhea1     amenorrhea2

Frequency|
Row Pct  |       0|       1|  Total
---------+--------+--------+
       0 |    285 |    104 |    389
         |  73.26 |  26.74 |
---------+--------+--------+
       1 |     31 |     56 |     87
         |  35.63 |  64.37 |
---------+--------+--------+
Total         316      160      476
```

Probabilities p_{00}, p_{01}, p_{10}, and p_{11} determine the distribution, where e.g.

$$p_{01} = P(Y_1 = 0 \text{ and } Y_2 = 1).$$

Describing dependence I: correlation

$$\mathrm{Cor}(Y_1, Y_2) = \frac{p_{11} - (p_{01} + p_{11})(p_{10} + p_{11})}{\sqrt{(p_{01} + p_{11})(p_{00} + p_{10})(p_{10} + p_{11})(p_{00} + p_{01})}}$$

- $\mathrm{Cor} = 0$ if Y_1 and Y_2 are independent.
- $\mathrm{Cor} = 1$ if all data are on the diagonal.
- $\mathrm{Cor} = -1$ if all data are off the diagonal.

Specifying the marginal probabilities and the correlation fully determines the bivariate distribution.

- **BUT:** The marginal probabilities impose restrictions on the correlation; ± 1 is only possible if they are identical.

Describing dependence II: conditional odds ratios

$$\mathrm{OR}(Y_2|Y_1) = \frac{\mathrm{odds}(Y_2 = 1|Y_1 = 1)}{\mathrm{odds}(Y_2 = 1|Y_1 = 0)} = \frac{p_{11}/p_{10}}{p_{01}/p_{00}}$$

- $\mathrm{OR} = 1$ if Y_1 and Y_2 are independent.
- $\mathrm{OR} > 1$ indicates a positive association.
- $\mathrm{OR} < 1$ indicates a negative association.

Specifying the marginal probabilities and the conditional odds ratio fully determines the joint distribution.

- with no restrictions on the conditional OR

BUT: Beware of empty cells: division by zero.

Amenorrhea: Pairwise associations

```
low dose correlations          high dose correlations

    1.00  0.40  0.28  0.27           1.00  0.31  0.25  0.29
    0.40  1.00  0.45  0.35           0.31  1.00  0.43  0.43
    0.28  0.45  1.00  0.53           0.25  0.43  1.00  0.47
    0.27  0.35  0.53  1.00           0.29  0.43  0.47  1.00
```

```
                  low dose                    high dose
Occations     cond.OR (95% CL)            cond.OR  (95% CL)
-------------------------------------------------------------------
(1,2)        8.16  (4.87, 13.66)         4.95 (3.02, 8.10)
(1,3)        4.52  (2.62, 7.800)         4.38 (2.36, 8.12)
(1,4)        5.03  (2.62, 9.660)         6.82 (3.12, 14.91)
(2,3)        9.29  (5.51, 15.69)         8.55 (5.01, 14.59)
(2,4)        6.22  (3.51, 11.00)         9.94 (5.36, 18.44)
(3,4)        12.58 (7.29, 21.70)         7.94 (4.91, 12.84)
```

Outline

Repetition: Binary data and logistic regression

Describing dependence between two binary outcomes

Population average models

Missing data in PA models

Subject specific models (GLMMs)

Comparison of PA and SS models

The PA model for repeated binary data

Marginal logistic regression to describe dependence on covaraites

$$\log\left\{\frac{P(Y_{ij}=1)}{1-P(Y_{ij}=1)}\right\} = \beta_0 + \beta_1 X_{1ij} + \ldots + \beta_k X_{kij}$$

Dependence between **pairs of replicates** are described by either

- Specifying a correlation matrix.
- Specifying a conditional odds ratio matrix .

Note: In case of more than two replicates this is only a partial model specification, but it is sufficient to estimate the regression parameters.

Interpretation of PA models

What effect does this covariate have **on the population**?

Typical applications are epidemiological studies where the interest is the prevalence of e.g. a disease in different subpopulations.

Data is clustered either by design or by nature, e.g.

- Children that are siblings.
- Patients belonging to the same clinic or GP.

and this must be accounted for in the statistical analyses.

Even: Subjects in a cohort followed over time if the main interest is how the population and not the individual changes with time.

Generalized Estimating Equations (technical)

Mimic generalized least squares estimation for linear mixed models; minimize weighted residual sum of squares, **sorry** matrix-notation,

$$\sum_{i=1}^{N}\{y_i - p_i(\beta)\}^T W_i^{-1}(\{y_i - p_i(\beta)\} = 0$$

W_i is the working covariance for subject i, given by

- The variances $\sigma_{ij}^2(\beta) = p_{ij}(\beta)\{1 - p_{ij}(\beta)\}$.
- The working correlation matrix C

Two steps are alternated until convergence:

- Minimize for β with the current estimates of W_i fixed.
- Estimate the correlation parameters from the standardized residuals $e_{ij} = \hat{\sigma}_{ij}^{-1}(Y_{ij} - \hat{p}_{ij})$ and update W_i.

The working correlation

- unstructured (`type=un`)
- autoregressive (`type=ar`)
- compound symmetry (`type=cs` or `type=exch`)
- working independence (`type=ind`)

Alternative option for modeling the conditional odds ratios.

- unstructured (`logor=fullclust`)
- two-level model (`logor=exch`) assuming exchangeability.
- similar but cluster dependent (`logor=logorvar(varname)`).
- three-level model (`logor=nest1 subclust=varname`)

GEE is most **efficient** when the model is close to the truth.

Properties of GEE

- The GEE-estimates are robust. They are consistent even when the working correlation C is misspecified (not the truth).
- The GEE-estimates are approximately normally distributed when sample size is large. This is used to construct confidence intervals and for hypothesis testing.
- Standard errors for the estimates should be estimated using the robust sandwich covariance estimator which is valid also when the working correlation is misspecified.

Statistical properties rely on having a large number of subjects!

Limitations of GEE

- Known to perform poorly in small datasets: you'll get anti-conservative SEs!
- No good for unbalanced longitudinal data. The sandwich estimator is useless when time points differ between subjects.
- When there are missing data that are MAR but not MCAR, the estimates may be biased.
- In case of missing data the GEE estimates should be computed using inverse probability weighting*, but this requires a LARGE sample size to work.

$\star$ see Fitzmaurice et al (2011), chapter 18

Amenorrhea: Modeling considerations

- We want to estimate the prevalences of amenorrhea at each time of follow-up and compare them between the two doses.
- The data is binary so a logistic model is appropriate. Covariates are dose and time (with interaction).
- There is no baseline measurement to deal with (no one suffered from amenorrhea before taking the drug).
- For the moment we **choose to ignore** the missing data.
- Dependence seem to decrease with increasing time space so an unstructured correlation is most appropriate.

PA model with proc genmod

```
proc genmod data=ameno descending;
class id dose time;
model amenorrhea=dose*time / noint dist=binomial;
repeated subject=id / withinsubject=time type=un corrw;
run;
```

- `descending`: models the probability that `amenorrhea=1`.
- `dist=binomial`: specifies the logistic model with `logit link`-function and the relevant variance w.r.t. the mean.
- `repeated` specifies the working correlation, here the `type=un`. The option `withinsubject=time` is important for getting the correct ordering of the correlated observations.
- `corrw` prints the estimated working covariance.

Alternative model specifications

We estimate risk in each group and at each time point separately:

```
model amenorrhea=dose*time / noint dist=binomial;
```

To test the interaction / compare changes in risk over time, use:

```
model amenorrhea=dose time dose*time / dist=binomial;
```

To fit the additive model (assuming no interaction) use:

```
model amenorrhea=dose time / dist=binomial;
```

Initial output

```
       Model Information

Data Set               WORK.AMENO
Distribution             Binomial
Link Function               Logit
Dependent Variable     amenorrhea


Number of Observations Read        4604
Number of Observations Used        3616
Number of Events                   1231
Number of Trials                   3616
Missing Values                      988


                ----

          Response Profile

 Ordered                          Total
   Value     amenorrhea       Frequency
       1     1                     1231
       2     0                     2385

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

Modelspecification and summary of data; Note the missing!

Select output: The working correlation

```
             GEE Model Information

Correlation Structure                   Unstructured
Within-Subject Effect                  time (4 levels)
Subject Effect                       id (1151 levels)
Number of Clusters                              1151
Clusters With Missing Values                     437
Correlation Matrix Dimension                       4
Maximum Cluster Size                               4
Minimum Cluster Size                               1


Algorithm converged.


               Working Correlation Matrix

             Col1         Col2         Col3         Col4

Row1       1.0000       0.3449       0.2614       0.2719
Row2       0.3449       1.0000       0.4367       0.3909
Row3       0.2614       0.4367       1.0000       0.5055
Row4       0.2719       0.3909       0.5055       1.0000
```

This describes the working model for the correlations.

Select output: Fixed effect estimates

```
                  Analysis Of GEE Parameter Estimates
                   Empirical Standard Error Estimates

                            Standard   95% Confidence
Parameter        Estimate     Error        Limits            Z Pr > |Z|

Intercept          0.0000    0.0000    0.0000    0.0000        .       .
dose*time 0 1     -1.4778    0.1071   -1.6878   -1.2678   -13.79   <.0001
dose*time 0 2     -1.0158    0.1020   -1.2157   -0.8160    -9.96   <.0001
dose*time 0 3     -0.4221    0.0989   -0.6160   -0.2283    -4.27   <.0001
dose*time 0 4      0.0596    0.1023   -0.1409    0.2601     0.58   0.5601
dose*time 1 1     -1.3540    0.1033   -1.5564   -1.1516   -13.11   <.0001
dose*time 1 2     -0.6400    0.0954   -0.8270   -0.4530    -6.71   <.0001
dose*time 1 3      0.0657    0.0998   -0.1298    0.2613     0.66   0.5099
dose*time 1 4      0.2614    0.1047    0.0562    0.4665     2.50   0.0125
```

Paramter estimates are the log-odds for amenorrhea in the two groups at the various occations. How do we back-transform to get the risk estimates?

Estimate statements

Compute risk estimates with.

```
estimate 'risk occation 1 high dose' dose*time 0 0 0 0 1 0 0 0;
estimate 'risk occation 2 high dose' dose*time 0 0 0 0 0 1 0 0;
estimate 'risk occation 3 high dose' dose*time 0 0 0 0 0 0 1 0;
estimate 'risk occation 4 high dose' dose*time 0 0 0 0 0 0 0 1;
estimate 'risk occation 1 low dose' dose*time 1 0 0 0 0 0 0 0;
estimate 'risk occation 2 low dose' dose*time 0 1 0 0 0 0 0 0;
estimate 'risk occation 3 low dose' dose*time 0 0 1 0 0 0 0 0;
estimate 'risk occation 4 low dose' dose*time 0 0 0 1 0 0 0 0;
```

Compute odds ratios for differences between groups (or occations):

```
estimate 'OR occation 1' dose*time -1 0 0 0 1 0 0 0 / exp;
estimate 'OR occation 2' dose*time 0 -1 0 0 0 1 0 0 / exp;
estimate 'OR occation 3' dose*time 0 0 -1 0 0 0 1 0 / exp;
estimate 'OR occation 4' dose*time 0 0 0 -1 0 0 0 1 / exp;
```

These statements **have to be added to the program** on slide XX.

Output from estimate statements

```
                        Contrast Estimate Results

                            Mean          Mean          L'Beta        L'Beta          Chi-
Label                    Estimate   Confidence Limits  Estimate   Confidence Limits  Square  Pr > ChiSq

risk occation 1 low        0.1858    0.1561    0.2196   -1.4778   -1.6878   -1.2678  190.26     <.0001
risk occation 2 low        0.2658    0.2287    0.3066   -1.0158   -1.2157   -0.8160   99.22     <.0001
risk occation 3 low        0.3960    0.3507    0.4432   -0.4221   -0.6160   -0.2283   18.23     <.0001
risk occation 4 low        0.5149    0.4648    0.5647    0.0596   -0.1409    0.2601    0.34     0.5601
risk occation 1 high       0.2052    0.1742    0.2402   -1.3540   -1.5564   -1.1516  171.94     <.0001
risk occation 2 high       0.3452    0.3043    0.3886   -0.6400   -0.8270   -0.4530   45.00     <.0001
risk occation 3 high       0.5164    0.4676    0.5649    0.0657   -0.1298    0.2613    0.43     0.5099
risk occation 4 high       0.5650    0.5140    0.6146    0.2614    0.0562    0.4665    6.23     0.0125

OR occation 1              0.5309    0.4581    0.6024    0.1238   -0.1679    0.4154    0.69     0.4055
Exp(OR occation 1)                                      1.1318    0.8455    1.5150
OR occation 2              0.5929    0.5255    0.6569    0.3758    0.1021    0.6495    7.24     0.0071
Exp(OR occation 2)                                      1.4562    1.1075    1.9147
OR occation 3              0.6196    0.5529    0.6820    0.4879    0.2126    0.7632   12.07     0.0005
Exp(OR occation 3)                                      1.6289    1.2369    2.1451
OR occation 4              0.5503    0.4787    0.6198    0.2018   -0.0851    0.4886    1.90     0.1680
Exp(OR occation 4)                                      1.2236    0.9184    1.6301
```

Here you find the estimated prevalences and the odds ratios comparing the groups.

What would happen if we ignored the repetitions?

(by leaving out the repeated statement).

Label	Independence	Unstructured correlation
risk 1 (high)	0.21 (0.17; 0.24)	0.21 (0.17; 0.24)
risk 2 (high)	0.34 (0.30; 0.38)	0.35 (0.30; 0.39)
risk 3 (high)	0.49 (0.44; 0.54)	0.52 (0.47; 0.56)
risk 4 (high)	0.54 (0.48; 0.59)	0.56 (0.51;0.61)
risk 1 (low)	0.19 (0.16; 0.22)	0.19 (0.16; 0.22)
risk 2 (low)	0.26 (0.22; 0.30)	0.27 (0.23; 0.31)
risk 3 (low)	0.39 (0.34; 0.44)	0.40 (0.35; 0.44)
risk 4 (low)	0.50 (0.45; 0.55)	0.51 (0.46; 0.56)
OR high vs low 1	1.13 (0.85; 1.51)	1.13 (0.85; 1.51)
OR high vs low 2	1.43 (1.08; 1.88)	1.46 (1.11; 1.91)
OR high vs low 3	1.53 (1.16; 2.03)	1.63 (1.24; 2.15)
OR high vs low 4	1.15 (0.85; 1.54)	1.22 (0.92; 1.63)

Nothing for occations 1, little for 2, etc - ?!? Why?

Within or betwen group differences

We are comparing prevalences at **each separate occation**

- Could have ignored the other occations.
- Then we have independent data!
- In fact, differences are **solely due to missing data**.

Estimated **changes over time** differ: here the pairing matters

```
                 assumed independence   unstructured correlation
----------------------------------------------------------------
OR 2 vs 1 (low)    1.56 (1.16;2.09)  vs   1.59 (1.26;2.00)
OR 3 vs 1 (low)    2.79 (2.09;3.72)  vs   2.87 (2.24;3.68)
OR 4 vs 1 (low)    4.41 (3.28;5.92)  vs   4.65 (3.60;6.01)
OR 2 vs 1 (high)   1.96 (1.49;2.59)  vs   2.04 (1.62;2.57)
OR 3 vs 1 (high)   3.77 (2.84;5.01)  vs   4.14 (3.23;5.30)
OR 4 vs 1 (high)   4.46 (3.34;5.97)  vs   5.03 (3.92;6.46)
```

The working correlation

Does our choice of working correlation affect the results?

1. working independence (may be inefficient)
2. unstructured correlation (as before).
3. conditional odds ratios (more natural interpretation)

```
Label                estimate1         estimate2           estimate3
risk 1 (high)  0.21 (0.17;0.24)  0.21 (0.17; 0.24)  0.21 (0.17; 0.24)
risk 2 (high)  0.34 (0.30;0.38)  0.35 (0.30; 0.39)  0.35 (0.30; 0.39)
risk 3 (high)  0.49 (0.44;0.54)  0.52 (0.47; 0.56)  0.52 (0.47; 0.57)
risk 4 (high)  0.54 (0.48;0.59)  0.56 (0.51; 0.61)  0.57 (0.52; 0.62)

risk 1 (low)   0.19 (0.16;0.22)  0.19 (0.16; 0.22)  0.19 (0.16; 0.22)
risk 2 (low)   0.26 (0.22;0.30)  0.27 (0.23; 0.31)  0.27 (0.23; 0.31)
risk 3 (low)   0.39 (0.34;0.44)  0.40 (0.35; 0.44)  0.40 (0.35; 0.44)
risk 4 (low)   0.50 (0.45;0.55)  0.51 (0.46; 0.56)  0.51 (0.46; 0.56)
```

The working correlation

```
 Odds ratio     working indep.   unstructured cor   conditional ORs
--------------------------------------------------------------
high vs low 1  1.13 (0.85;1.51)  1.13 (0.85;1.51)  1.13 (0.85;1.51)
high vs low 2  1.43 (1.08;1.88)  1.46 (1.11;1.91)  1.46 (1.11;1.92)
high vs low 3  1.53 (1.16;2.03)  1.63 (1.24;2.15)  1.63 (1.24;2.15)
high vs low 4  1.15 (0.85;1.54)  1.22 (0.92;1.63)  1.23 (0.92;1.64)

2 vs 1 (low)  1.56 (1.23;1.97)  1.59 (1.26;2.00)  1.59 (1.26;2.00)
3 vs 1 (low)  2.79 (2.16;3.59)  2.87 (2.24;3.68)  2.88 (2.25;3.69)
4 vs 1 (low)  4.41 (3.39;5.72)  4.65 (3.60;6.01)  4.65 (3.60;6.00)

2 vs 1 (high) 1.96 (1.55;2.49)  2.04 (1.62;2.57)  2.05 (1.62;2.58)
3 vs 1 (high) 3.77 (2.93;4.87)  4.14 (3.23;5.30)  4.15 (3.24;5.32)
4 vs 1 (high) 4.46 (3.44;5.79)  5.03 (3.92;6.46)  5.06 (3.94;6.49)
```

Comments

- Working indpendence is not the same as assuming independence (since the sandwich covariance estimator accounts for correlation).
- We get higher efficiency (more accurate estimates) when choosing a working correlation that is close to the truth; Working independence is usually not very efficient.
- We don't see much difference in results with working correlations of similar complexity.
- Simple correlation patterns such as CS or AR may often simplify computation without loosing much efficiency.

Pros and cons of the PA model

Advantages

- Direct estimation of population risk.
- Minimal model assumptions, hence fewer wrong assumptions.
- Computationally simple.

Drawbacks

- Poor small sample performance (anti-conservative SEs).
- Need additional modeling to handle missing data even when they are MAR and this is technical.
- Only pairwise dependency is modeled, no full multivariate model specification to use for e.g. power calculations.
- Possible to specify mean-correlation combinations that do not match any real distribution (conditional odds ratios are ok).

Outline

Repetition: Binary data and logistic regression

Describing dependence between two binary outcomes

Population average models

Missing data in PA models

Subject specific models (GLMMs)

Comparison of PA and SS models

Missing data

Missing data should not be ignored as the GEE-estimates may become biased.

We have a problem assessing prevalences of amenorrhea:

- Replicates are correlated; some women are more prone to experience amenorrhea than others.
- Women with amenorrhea at the previous occation are more likely to drop out.
- Then women with amenorrhea are underrepresented at later occations.
- Even more of a problem if inclination to drop out is differential between the treatment groups.

Amenorrhea: drop out

How does drop out depend on dose and previous response?

```
Time1    dose     amenorrhea1    N       drop out
-----------------------------------------------------------
           0                0    469       0.16
                            1    107       0.21
           1                0    457       0.15
                            1    118       0.26
------------------------------------------------------------
```

```
Time2    dose   amenorrhea1  amenorrhea2    N   drop out
-------------------------------------------------------------
         0             0            0     322    0.13
                                    1      71    0.20
                       1            0      30    0.10
                                    1      54    0.15
         1             0            0     285    0.14
                                    1     104    0.26
                       1            0      31    0.19
                                    1      56    0.27
```

Inverse probability weighting

Solution:

- Assign more weight to those who should be more inclined to drop out but actually stays in the trial.
- E.g. if women with amennorhea are twice as likely to drop out than those without, those who stay in has to count for two.

Inverse probability weighting can handle data that are **MAR**.

1. Model the probabilities of **not dropping out** as depending on dose and previous responses, e.g. in logistic regression models.
2. Input the inverse probability weights to `proc genmod`* to adjust for the missing data.

How to do this is described in Fitzmaurice et al (2011), chapter 18.

Estimates with missing data

```
                   unstructured cor    IPW-estimates     complete cases
-----------------------------------------------------------------------
risk 1 (low)       0.19 (0.16;0.22)  0.19 (0.16;0.22)  0.18 (0.14;0.22)
risk 2 (low)       0.27 (0.23;0.31)  0.27 (0.23;0.31)  0.25 (0.21;0.30)
risk 3 (low)       0.40 (0.35;0.44)  0.40 (0.35;0.44)  0.37 (0.32;0.42)
risk 4 (low)       0.51 (0.46;0.56)  0.52 (0.46;0.57)  0.50 (0.45;0.55)

risk 1 (high)      0.21 (0.17;0.24)  0.21 (0.17;0.24)  0.16 (0.13;0.20)
risk 2 (high)      0.35 (0.30;0.39)  0.34 (0.30;0.39)  0.30 (0.25;0.35)
risk 3 (high)      0.52 (0.47;0.56)  0.52 (0.47;0.57)  0.48 (0.43;0.53)
risk 4 (high)      0.56 (0.51;0.61)  0.57 (0.52;0.62)  0.54 (0.48;0.59)

high vs low (1)  1.13 (0.85;1.51) 1.13 (0.85;1.51) 0.89 (0.60;1.32)
high vs low (2)  1.46 (1.11;1.91) 1.45 (1.09;1.91) 1.24 (0.89;1.72)
high vs low (3)  1.63 (1.24;2.15) 1.63 (1.23;2.16) 1.56 (1.15;2.10)
high vs low (4)  1.22 (0.92;1.63) 1.23 (0.92;1.65) 1.15 (0.85;1.54)
```

Complete case analysis should not be performed - it is biased.

The MAR assumption

The MAR assumption can never be assessed from the data!

Why do subjects drop out?

- **Ask them!**
- Include as much background information as possible in a non-response analysis.

What you need to **argue**:

- MAR means that there are no unmeasured factors that are predictive of both response and drop out.

Outline

Repetition: Binary data and logistic regression

Describing dependence between two binary outcomes

Population average models

Missing data in PA models

Subject specific models (GLMMs)

Comparison of PA and SS models

Generalized linear mixed models

In parallel to a variance component models, we specify a model for the probability of the outcome of interest

$$\log\left\{\frac{P(Y_{ij}=1)}{1-P(Y_{ij}=1)}\right\} = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + b_i$$

including a random intercept (or other random effects).

Model assumption: b_i's are normally distributed

Note: This fully specifies a multivariate model.

- The random intercept induce correlation within subejcts.

Interpretation of variance components

Variability among subject is modeled by the random intercept b_i.

- Can think of these as varying susceptabilities.

Random effects are **assumed to follow a normal distribution**,

- It is difficult to check the validity of this assumption, but fortunately inference for the regression parameters is **not very sensitive** to misspecification of the distribution.
- We could use the normal range for the b_i's to visualize variation in e.g. risk among subjects, but be aware that this is an **extrapolation** strongly influenced by the assumption.
- The random effects can be estimated by the BLUPs, but again these estimates are also strongly influenced by the normal assumption and thus no good for checking it!

Interpretion of regression parameters

Regression parameters have subject specific interpretations.

- The effect of increasing X by one unit while keeping all other covariates **including the random intercept** b fixed.

E.g. the effect of treatment on the individual.

- Most reasonable if treatment is a within subject covariate (or at least randomized so that groups have similar b's).
- No sensible subject specific interpretation of non-randomized covariates such as gender - !

Typical application is longitidunal data:

- Interest is how the individual risk changes with time.

Estimation in GLMMs (technical)

Maximize the likelihood function:

$$L(\beta, G) = \prod_{i=1}^{N} \int L_{\text{binary}}(y_i|\beta, b_i) L_{\text{normal}}(b_i|G) db_i$$

- a *mixture* of the logistic likelihood for the binary outcome w.r.t. the normal likelihood for the variance component.

In practice estimation has to rely on **numerical integration**.

- We recommend Gaussian quadrature with a suitably large number of quadrature points.
- Computation may take several minutes (or even hours).
- Infeasible with many random effects or if data is very large.

Amenorrhea: modeling considerations

- We want to estimate how individual risk of amenorrhea changes with duration of the drug.
- We use a logistic GLMM for binary data and with dose and time as fixed effects (including the interaction).
- We include a random intercept to model between subjects variation in susceptability.
- A random regression type model which further includes a random slope might be a more realistic model for longitudinal data, but with only four time points it is a bit overkill.
- We need not worry about the missing data because GLMMs automatically adjusts for data that are MAR.

PROC GLIMMIX

```
proc glimmix data=ameno method=quad(qpoints=50) noclprint;
class id dose time;
model amenorrhea = dose*time / noint dist=binomial solution;
random intercept / subject=id;
run;
```

- `method=quad`: approximates the likelihood function by Gaussian quadrature.
- `qpoints=50`: the more quadrature points, the better accuracy
- `random`: here we have only a single random effect, so it is not necessary to specify `type=un` and to print the g-matrix.

Initial output

```
The GLIMMIX Procedure

                  Model Information

Data Set                     WORK.AMENO
Response Variable            amenorrhea
Response Distribution        Binomial
Link Function                Logit
Variance Function            Default
Variance Matrix Blocked By   id
Estimation Technique         Maximum Likelihood
Likelihood Approximation     Gauss-Hermite Quadrature
Degrees of Freedom Method    Containment


Number of Observations Read        4604
Number of Observations Used        3616


               Dimensions

G-side Cov. Parameters           1
Columns in X                     8
Columns in Z per Subject         1
Subjects (Blocks in V)        1151
Max Obs per Subject              4
```

Output: Numerical optimisation

```
           Optimization Information

Optimization Technique        Dual Quasi-Newton
Parameters in Optimization    9
Lower Boundaries              1
Upper Boundaries              0
Fixed Effects                 Not Profiled
Starting From                 GLM estimates
Quadrature Points             50

                               Iteration History

                                              Objective                        Max
Iteration    Restarts    Evaluations           Function          Change   Gradient
        0           0              4        3959.248351           .       72.35469
        1           0              2       3901.1681151     58.08023598   27.62877
        2           0              2       3882.5814401     18.58667495   29.26711
        3           0              3       3871.0423322     11.53910788   16.05843
        4           0              3       3865.1824909      5.85984136   6.835517
        5           0              3       3864.7577602      0.42473070   2.875723
        6           0              2       3864.0629681      0.69479210   2.356003
        7           0              3       3863.8737183      0.18924976    1.13089
        8           0              3        3863.818349      0.05536933   0.907351
        9           0              3       3863.8098004      0.00854858   0.556536
       10           0              3       3863.8050711      0.00472927   0.121306
       11           0              3       3863.8047589      0.00031221   0.023196
       12           0              3       3863.8047557      0.00000325   0.011083

         Convergence criterion (GCONV=1E-8) satisfied.
```

Output: parameter estimates

```
        Covariance Parameter Estimates

                                       Standard
Cov Parm     Subject     Estimate         Error
Intercept    id            5.1011        0.5873


                         Solutions for Fixed Effects

                                             Standard
Effect       dose     time     Estimate         Error       DF    t Value    Pr > |t|

dose*time    0        1         -2.5581        0.1903     2459     -13.45      <.0001
dose*time    0        2         -1.7877        0.1820     2459      -9.82      <.0001
dose*time    0        3         -0.7605        0.1760     2459      -4.32      <.0001
dose*time    0        4         0.08815        0.1805     2459       0.49      0.6253
dose*time    1        1         -2.3492        0.1832     2459     -12.82      <.0001
dose*time    1        2         -1.1409        0.1690     2459      -6.75      <.0001
dose*time    1        3          0.1129        0.1742     2459       0.65      0.5169
dose*time    1        4          0.4590        0.1819     2459       2.52      0.0117
```

Estimated log-odds for a subject of average susceptibility.

Estimates: changes over time

Add the following to the program on slide XX:

```
estimate 'OR 2 vs 1 (l)' dose*time -1 1 0 0 0 0 0 0/ exp cl;
estimate 'OR 3 vs 1 (l)' dose*time -1 0 1 0 0 0 0 0/ exp cl;
estimate 'OR 4 vs 1 (l)' dose*time -1 0 0 1 0 0 0 0/ exp cl;
estimate 'OR 2 vs 1 (h)' dose*time 0 0 0 0 -1 1 0 0/ exp cl;
estimate 'OR 3 vs 1 (h)' dose*time 0 0 0 0 -1 0 1 0/ exp cl;
estimate 'OR 4 vs 1 (h)' dose*time 0 0 0 0 -1 0 0 1/ exp cl;
```

```
Odds ratio (95% CL)       Low dose             High dose
-------------------------------------------------------------
Occation 2 vs 1        2.16 (1.46;3.19)      3.35 (2.29;4.89)
Occation 3 vs 1        6.04 (4.01;9.08)     11.73 (7.67;17.93)
Occation 4 vs 1       14.10 (9.07;21.93)    16.58 (10.61;25.91)
```

Substantial increase in **individual risk** over time.

Pros and cons of GLMMs

Advantages

- Can model more complex variance component structures.
- Fully specified model allows for exact likelihood inference, model comparisons, and power calculations.
- Likelihood inference automatically handles MAR optimally.

Drawbacks

- No estimates of population risk (and/or diffrences)
- More model assumptions, thus higher risk of misspecification.
- **Impossible to check assumptions about the random effects** and very tempting to extrapolate from these.
- Computationally infeasible when the number of random effects or the overall size of the data becomes large.

Outline

Repetition: Binary data and logistic regression

Describing dependence between two binary outcomes

Population average models

Missing data in PA models

Subject specific models (GLMMs)

Comparison of PA and SS models

Differences beween PA and SS models

Conceptual differences

- Population risk vs individual risk.
- Choose a model that answers your scientific questions.

Modeling differences

- Covariance pattern or variance component model?
- GEE is more robust w.r.t. model violations.
- Choose a model that is realistic for your data.

Computational differences

- Handling of missing data (MAR) easier in GLMMs.
- GLMMs become infeasible with large data.
- GEE-standard errors are biased in small data.
- Choose a model that can handle your data.

Population vs individual risk

Anticipated risk of a subject with average susceptibitiy is not the same as the population risk or prevalence!

Esitmated risks in the Amenorrhea study:

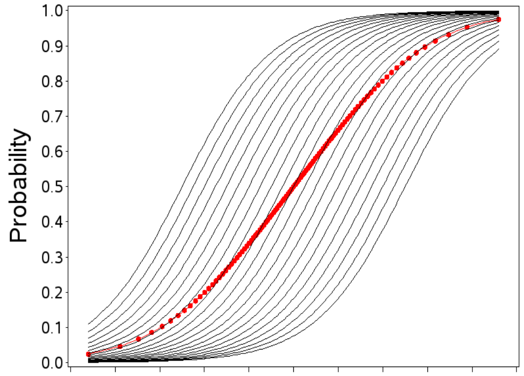
```
                  PA (with IPWs)     SS (random intercept)
risk 1 (low)      0.19 (0.16;0.22)      0.07 (0.05;0.10)
risk 2 (low)      0.27 (0.23;0.31)      0.14 (0.10;0.19)
risk 3 (low)      0.40 (0.35;0.44)      0.32 (0.25;0.40)
risk 4 (low)      0.52 (0.46;0.57)      0.52 (0.43;0.61)

risk 1 (high)     0.21 (0.17;0.24)      0.09 (0.06;0.12)
risk 2 (high)     0.34 (0.30;0.39)      0.24 (0.19;0.31)
risk 3 (high)     0.52 (0.47;0.57)      0.53 (0.44;0.61)
risk 4 (high)     0.57 (0.52;0.62)      0.61 (0.53;0.69)
```

Estimates in the SS model are always further away from 0.50 compared to the PA model, but on the same side.

Hypothetical example for illustration

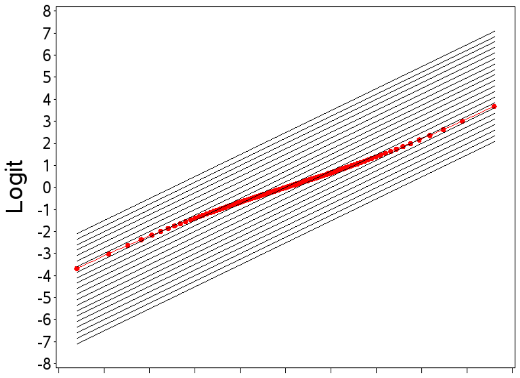
Subject specific model with a covariate effect (x-axis) and 21 individual curves.



The population average follows a less steep curve and moreover this does not have the functional form of a logistic model . . .

Population average on logit scale

The subject specific model yields parallel lines on logit scale



but the population average deviates from a straight line – and has a less steep slope (smaller effect of covariate x)

Odds ratios in PA and SS models

Odds ratios from the SS model are always further away from 1 than in the PA model, but on the same side.

Odds ratios in amenorrhea study:

```
                PA (with IPWs)         SS (random intercept)
h vs l (1)  1.13 (0.85-1.51, P=0.41)  1.23 (0.77-1.98, P=0.39)
h vs l (2)  1.45 (1.09-1.91, P=0.01)  1.91 (1.19-3.06, P=0.01)
h vs l (3)  1.63 (1.23-2.16, P=0.00)  2.40 (1.47-3.90, P=0.00)
h vs l (4)  1.23 (0.92-1.65, P=0.17)  1.45 (0.88-2.40, P=0.15)

2 vs 1 (l)  1.59 (1.26-2.01, P=0.00)  2.16 (1.46;3.19, P=0.00)
3 vs 1 (l)  2.88 (2.24-3.71, P=0.00)  6.04 (4.01;9.08, P=0.00)
4 vs 1 (l)  4.67 (3.60-6.06, P=0.00)  14.1 (9.07;21.9, P=0.00)

2 vs 1 (h)  2.04 (1.61-2.57, P=0.00)  3.35 (2.29-4.89, P=0.00)
3 vs 1 (h)  4.15 (3.23-5.34, P=0.00)  11.7 (7.67-17.9, P=0.00)
4 vs 1 (h)  5.08 (3.93-6.56, P=0.00)  16.6 (10.6-25.9, P=0.00)
```

Note: Very similar p-values.

Linear and non-linear models

These findings are **contrary to the linear mixed models**.

- The variance component model with random intercept

  `random intercept / subject=id;`

- The covariance pattern model with compound symmetry

  `repeated time / subject=id type=cs;`

. . . are exactly the same!

Concluding remarks

PA or SS models – which should be preferred?

- Different specifications of the joint distribution of Y_i lead to regression coefficients with quite distinct interpretations.
- PA models aim at inference for the population means. The models are only partially specified. They merely acknowledge the correlation and do not seek to explain it.
- SS models assume that the correlation among repeated measurements arise from the sharing of random effect or subject specific parameters. The parameters in GLMMs have subject specific but not population average interpretations.

Fitzmaurice et al (2011) says:

> *Choice of model should be made on subject matter grounds. There is no contradiction in repporting both subject-specific and population averaged effects if both are of interests.*