An Introduction to Generalized Linear Mixed Models

by
Charles E. McCulloch
Division of Biostatistics
Department of Epidemiology and Biostatistics
University of California, San Francisco

Outline

1) Introduction
   a) Example: Back pain
   b) Overview of workshop
   c) Hierarchical modeling
2) Review: Linear Mixed Models (LMMs)
   a) Example: Fecal fat
   b) Example: Propranolol
      i) Analysis
      ii) Correlations
      iii) Shrinkage estimators
   c) Fixed versus random factors
   d) Best Linear Unbiased Prediction
   e) Estimation, tests and software in LMMs
3) Review: Generalized Linear Models (GLMs)
   a) Example: Potato flour dilutions
   b) GLM Basics
   c) Example: snap beans
   d) Transforming versus linking
   e) Estimation, tests and software for GLMs
4) Introduction to GLMMs
   a) Example: skin cancer
5) Modeling in GLMMs

      a)   Progabide and seizures
      b)   Cartoons and learning disabilities
      c)   Photosynthesis in corn
      d)   Chestnut leaf blight
      e)   Combat vehicle design
      f)   Troponin and cardiac damage

6) Features of GLMMs
      a)   Consequences of model assumptions
      b)   Marginal versus conditional models

7) Inference for GLMMs
      a)   Maximum likelihood
      b)   Conditional inference
      c)   Generalized estimating equations
      d)   Penalized quasi-likelihood
      e)   Best Prediction for GLMMs
      f)   Software

8) Case Studies
      a)   Breeding Bird Survey
      b)   Progabide and seizures
      c)   Potomac River Fever in horses
      d)   Chestnut Leaf Blight

9) Summary/Discussion

Approximate Schedule

Morning: 8:30-12:00, Lunch 12:00-13:15, Afternoon: 13:15-14:00

Monday

    <u>Morning</u>: Introduction, review of linear mixed models (LMMs) and generalized linear models (GLMs)
    <u>Afternoon</u>: Introduction to generalized linear mixed models (GLMMs); GLMM modeling.

Tuesday

    <u>Morning</u>: GLMM modeling (cont), features of GLMMs, inference for GLMMs.
    <u>Afternoon</u>: Case studies and Summary/Discussion

References

Abramowitz, M. and Stegun, I.A. (1964). Handbook of Mathematical Functions. National Bureau of Standards, Washington, D.C.

Abu-Libdeh, H., Turnbull, B. and Clark, L.C. (1990) Analysis of multi-type recurrent events in longitudinal studies: Application to a skin cancer prevention trial. *Biometrics* **46**: 1017-1034.

Aitken, M. (1999). A general maximum likelihood analysis of variance components. *Biometrics* **55**: 117-128

Atwill, E.R., H.O. Mohammed, J.W. Lopez, C.E. McCulloch, and E.J. Dubovi. Cross-sectional evaluation of environmental, host, and management factors associated with the risk of seropositivity to *Ehrlichia risticii* in horses of New York State. *American Journal of Veterinary Research*, 57: 278-285, 1996.

Borsch-Supan, A. and Hajivassiliou, V. (1993). Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variables. *Journal of Econometrics* **58**: 347-368.

Booth, J., and Hobert, J. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association* **93**: 262-272.

Booth, J.G. and Hobert, J.H. (1999), "Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm", *Journal of the Royal Statistical Society* B 62: 265-285.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**: 9-25.

Breslow, N.E. and Lin, X. (1994). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika* **82**:81-91.

Carey, V., Zeger, S.L. and Diggle, P. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika* **80**: 517-526.

Casella, G.C. and Berger, R.L. (1994). Estimation with selected binomial information, or Do you really believe Dave Winfield is batting .471? *Journal of the American Statistical Association* **89**: 1080-1090.

Conaway, M.R. (1989). Analysis of repeated categorical measurements with conditional likelihood methods. *Journal of the American Statistical Association* **89**: 53-62.

Conaway, M.R. (1990). A random effects model for binary data. *Biometrics* **46**: 317-328.

Cortesi, P., and Milgroom, M. (1998). Genetics of vegetative incompatibility in *Cryphonectria parasitica Appl. Environ Microbiol.* **64**: 2988-2994.

Cortesi, P., Milgroom, M., and Bisiach, M. (1996). Distribution and diversity of vegetative incompatibility types in subpopulations of *Cryphonectria parasitica* in Italy. *Mycological Research*, **100**: 1087-1093.

Cortesi, P., McCulloch, C, Song, H., Lin, H. and Milgroom, M. (2001). Genetic control of horizontal virus transmission in the chestnut blight fungus, *Cryphonectria parasitica*. *Genetics*, **159**: 107-118.

Cox, D.R. and Snell E.J. (1989). *Analysis of Binary Data*, 2*nd* Edition. Chapman and Hall, London.

Crowder, M.J. (1978). Beta-binomial ANOVA for proportions. *Applied Statistics* **27**:34-37.

Devore, J. and Peck, R. (1993). Statistics. Duxbury, Belmont, CA.

Diggle, P., Liang, K.-Y., Zeger, S.L. and Heagerty, P (2002). Longitudinal Data Analysis, 2nd Ed. Oxford University Press, Oxford.

Drum, M. and McCullagh, P. (1993). REML estimation with exact covariance in the logistic mixed model. *Biometrics* **49**: 677-689.

Engel, B. and Keen, A. (1994). A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica* **48**:1-22.

Fitzmaurice, G.M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics* **51**: 309-317.

Geyer, C.J. (1994). Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. Technical Report No. 568, School of Statistics, University of Minnesota.

Geyer, C.J. and Thompson, E.A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society, Series B,* **54**: 657-699.

Gilks, W.R., Wang, C.C., Yvonnet, B., and Coursaget, P. (1993). Random-effects models for longitudinal data using Gibbs sampling. *Biometrics* **49**: 441-453.

Gilmour, A.R., Anderson, R.D. and Rae, A.L. (1984). The analysis of binary data by a generalized linear mixed model. *Biometrika* **72**: 593-599.

Hamet, P., Kuchel, O., Cuche, J.L., Boucher, R., and Genest, J. (1973). Effect of propranolol on cyclic AMP excretion and plasma renin activity in labile essential hypertension. *Canadian Medical Association Journal* **1**: 1099-1103.

Heagerty, P. and Lele, S. (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association* **93**: 1099-1111.

Heagerty, P. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics* **55**: 688-698.

Heagerty, P. and Kurland, B. (2001). Misspecified maximum likelihood estimates and generalized linear mixed models. *Biometriak* **88**: 973-986.

Hedeker, D. and Gibbons, R.D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics* **50**: 933-944.

James, F.C., McCulloch, C.E., and Wiedenfeld, D.A. (1996). New approaches to the analysis of population trends in land birds. *Ecology* **77**: 13-27.

Karim, M.R., Zeger, S.L. (1992). Generalized linear models with random effects; Salamander mating revisited. *Biometrics* **48**: 631-644.

Kauermann, G. and Carroll, R.J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *J. Amer. Stat. Association* **79**: 1387-1396.

Korff, M.V., Barlow, W., Cherkin, D., and Deyo, R.A. (1994). Effects of practice style in managing back pain. *Ann. Internal Med.* **121**: 187-95.

Kuk, A.Y.C. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society, Series B* **57**:395-407.

Lee, Y., and Nelder, J.A. (1996). Hierarchical generalized linear models (Disc: p656-678). *Journal of the Royal Statistical Society, Series B,* **58**: 619-656.

Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**:13-22

Liang, K.-Y., Zeger, S.L., and Qaqish, B.H. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B,* **54**: 673-687.

Lin, X. and Breslow, N.E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association* **91**: 1007-1016.

Lindsey, J.K. and Lambert, P. (1998). On the appropriateness of marginal models for repeated measures in clinical trials. *Statistics in Medicine* **17**: 447-469.

Lipsitz, S.R., Fitzmaurice, G.M., Orav, E.J., and Laird, N.M. (1994). Performance of generalized estimating equations in practical situations. *Biometrics* **50**: 270-278.

McCullagh, P. and Nelder, J. (1989). Generalized Linear Models, 2nd Ed. Chapman and Hall, London.

McCulloch, C.E. (1994). Maximum likelihood estimation of variance components for binary data. *Journal of the American Statistical Association* **89**: 330-335.

McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* **92**: 162-170.

McCulloch, C.E. and Searle, S.R. (2000). Generalized, Linear, and Mixed Models. Wiley, New York.

McGilchrist, C. A. (1993). REML estimation for survival models with frailty. *Biometrics* **49**: 221-225.

McGilchrist, C.A. (1994). Estimation in generalized mixed models. *Journal of the Royal Statistical Society, Series B* **56**: 61-69.

McGilchrist, C.A. and Yau, K. K. W. (1995). The derivation of BLUP, ML, REML estimation methods for generalised linear mixed models. *Communications in Statistics. Theory and Methods* **24**: 2963-2980.

McLean, R.A., Sanders, W.L., and Stroup, W.W. (1991). A unified approach to mixed linear models. *American Statistician* **45**: 54-64.

Natarajan, R. and McCulloch, C.E. (1995). A note on existence of the posterior distribution for a class of mixed models for binomial responses. *Biometrika* **82**:639-643.

Neuhaus, J. M., Lesperance, M. L. (1996). Estimation efficiency in a binary mixed-effects model setting. *Biometrika* **83**: 441-446.

Prentice, R.L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**: 1033-1048.

Robinson, G.K. (1991). That BLUP is a good thing - the estimation of random effects. *Statistical Sciences* **6**:15-51.

Ruppert, D., Cressie, N., and Carroll, R. (1989). A transformation/weighting model for estimating Michaelis-Menten parameters. *Biometrics* **45**: 637-656.

Ruppert, D., Reish, R.L., Deriso, R.B., and Carroll, R.J. (1984). Optimization using stochastic approximation and Monte Carlo simulation (with application to harvesting of Atlantic Menhaden). *Biometrics* **40**: 535-545.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* **78**:719-727.

Self and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82**: 605-610.

Searle, S.R., Casella, G., and McCulloch, C.E. (1992). Variance Components. Wiley, New York.

Snedecor, G.W. and Cochran, W.G. (1989). Statistical Methods, 8*th* Edition. Iowa State University Press, Ames, Iowa.

Stram, D.O. and Lee, J.W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* **50**:1171-1177.

Waclawiw, M.A., and Liang, K.-Y. (1993). Prediction of random effects in the generalized linear model. *Journal of the American Statistical Association* **88**: 171-178.

Wolfinger, R.W. (1994). Laplace's approximation for nonlinear mixed models. *Biometrika* **80**: 791-795.

Zeger, S.L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**: 121-130.

Zeger, S.L. and Karim, M.R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association* **86**: 79-86.

1. Introduction

Example: Practice style and back pain (Korff, Barlow, Cherkin, and Deyo, 1994).

Forty-four primary care physicians in a large HMO were classified according to their practice style in treating back pain management (low, moderate or high frequency of prescription of pain medication and bed rest). An average of 24 patients per physician was followed for 2 years (1 month, 1 year and 2 year followups) after the indexed visit. Outcome measures included functional measures (pain intensity, activity limitation days, etc.), patient satisfaction (e.g., "After your visit with the doctor, you fully understood how to take care of your back problem"), and cost.

Q1. Does practice style influence function, satisfaction or cost?

Q2. How much of the variability in the responses is due to physician?

Q3. How well are individual physicians performing with regard to effectiveness and cost?

Model for log(cost) at year 1: Incorporating predictors of practice style of the physician, age of the patient, whether the back pain was cervical or thoracic or neither (yes=1, no=0).

Model for "understand how to care for your back" at year 1: Incorporating predictors of practice style of the physician, age of the patient, whether the back pain was cervical or thoracic or neither (yes=1, no=0).

1. b. Overview

- Hierarchical modelling
- Review of Linear Mixed Models (LMMs) and Generalized Linear Models (GLMs)
- Examples of Generalized Linear Mixed Models (GLMMs)
- Modeling using GLMMs
- Features of GLMMs
- Inference methods
- Case studies
- Discussion and summary

1. c. Hierarchical Modeling

Hierarchical data: Data (responses and/or predictors) collected from different levels within a study. Other terminology for the same or related ideas: repeated measures data, longitudinal data, clustered data, multilevel data.

Example 1: Practice style and back pain (Korff, Barlow, Cherkin, and Deyo, 1994).

Example 2: The Educational Testing Service in the past has offered guidance to both law school admissions officers and to potential applicants to law school via their Law School Validity Studies. One aspect of this has been to create a simple index that allows admission officers to screen applicants and for applicants to gauge the likelihood of acceptance to a law school before applying.

Two of the indicators used for predicting success in law school are the LSAT score (ranging from 200 to 800) and the undergraduate GPA (UGPA). A form of combining the LSAT and UGPA, which has successfully been used in the past to predict first year performance at law school, has been:

Predicted performance = LSAT + (mult)×UGPA

Where mult is a multiplier chosen to reflect the relative importance of LSAT and UGPA and which might be dependent on the school doing the admissions. For example, a multiplier of 200 might make sense since it puts both GPA

(typically in a range of 1.0 to 4.0) and LSAT on the same scale.

In practice the multipliers have been estimated from data taken from admitted students and this is done separately for each law school. The estimation was often done by a multiple regression of first year performance on both LSAT and UGPA.

Q. What is the best way to estimate the multipliers for each school?

Example 3: Lack of digestive enzymes in the intestine can cause bowel absorption problems. This will be indicated by excess fat in the feces. Pancreatic enzyme supplements can be given to ameliorate the problem. Does the supplement form make a difference? (Graham DY, Enzyme replacement therapy of exocrine pancreatic insufficiency in man. *NEJM*, **296**: 1314-17, 1977 – But note: sex information made up for illustration.)

PatID / Sex	Fecal Fat (g/day) Pill type				Avg
	None	Tablet	Capsule	Coated Capsule	
1 – M	44.5	7.3	3.4	12.4	16.900
2 – M	33.0	21.0	23.1	25.4	25.625
3 – M	19.1	5.0	11.8	22.0	14.475
4 – F	9.4	4.6	4.6	5.8	6.100
5 – F	71.3	23.3	25.6	68.2	47.100
6 – F	51.2	38.0	36.0	52.6	44.450
Avg	38.08	16.53	17.42	31.07	25.775

Example 4: Propranolol and hypertension (Hamet, *et al*, 1975)

Below are data from an early, double-blind trial of the effect of propranolol on labile hypertension. Blood pressure was measured under the drug and a placebo both in the upright and recumbent positions.

| | Blood Pressure (mmHg) | | | | |
| | Recumbent | | Upright | | |
Patient	Placebo	Propran.	Placebo	Propran	Ave.
1	96	71	73	87	81.75
2	96	85	104	76	90.25
3	92	89	83	90	88.50
4	97	110	101	85	98.25
5	104	85	112	94	98.75
6	100	73	101	93	91.75
7	93	81	88	85	86.75
Ave.	96.86	84.86	94.57	87.14	90.86

Q1: Does Propranolol have the same influence in recumbent and upright positions?
Q2: If the answer to Q1 is yes, is it effective?

Analysis Approaches

Basic tenet: Don't go beyond standard and accepted statistical practices unless necessary.

Applied in this context: Do we need hierarchical models?

The usual statistical methods (multiple regression, basic ANOVA, logistic regression, and many others) assume observations are independent.

Important idea: observations taken within the same subgroup in a hierarchy are often more similar to one another than to observations in different subgroups, other things being equal. [correlated]

Also getting the correlation assumptions wrong in a statistical analysis is often a very serious mistake.

Simple Analysis Strategies

What strategies might we employ in analyzing data from a hierarchical format?

1. Separate analyses for each subgroup.
2. Analyses at the lowest level in the hierarchy.
3. Analyses at the highest level in the hierarchy.
4. Derived variables.

Let's consider an example of each of these and advantages and disadvantages.

1. <u>Separate analyses for each subgroup.</u>

Fecal fat example?

The law school example follows this approach by calculating separate multipliers for each law school.   Here are the multipliers estimated for selected law schools for three consecutive years and pooling the data across years.

Law School	Separate Years			Pooled Years	
	Year 1	Year 2	Year 3	Years 1-2	Years 2-3
1	2507	301	105	526	164
2	-24	49	153	5	116
3	179	118	98	149	107

Law schools 1 and 2 were selected as being somewhat extreme and 3 was "typical".

2. <u>Analyses at the lowest level in the hierarchy</u>.

For the back pain example this corresponds to analyzing each observation on the patient and attributing to each one the higher level characteristics, e.g., an observation taken from a "low" doctor.

3. <u>Analyses at the highest level in the hierarchy</u>.

For the back pain example, this corresponds to calculating the average value of the response for each doctor across all patients and time periods.

4. <u>Derived variable approach</u>.

For the fecal fat example we would calculate several new responses: (1) the difference between the none and tablet observations for each patient, (2) the difference between the none and capsule observations for each patient, (3) the difference between the and tablet and coated tablet observations for each patient. These new responses are then subjected to one-sample t-tests.

When to Use Hierarchical Models

The use of hierarchical/mixed models is clearly indicated in three situations:

1. When the correlation structure is of primary interest.

2. When we wish to "borrow strength" across the levels of a hierarchy in order to improve estimates.

   (81 law schools and one year of data versus 2 years of data)

3. When dealing with highly unbalanced correlated data.

2. Review: Linear Mixed Models (LMMs)

Analysis of the fecal fat example (Stata)

```
summ

Variable |      Obs        Mean    Std. Dev.       Min        Max
---------+-----------------------------------------------------------
  fecfat |       24      25.775    20.00214        3.4       71.3
   patid |       24         3.5    1.744557          1          6
pilltype |       24         2.5     1.14208          1          4

. sort pilltype

. by pilltype: summarize fecfat

-> pilltype=    none
Variable |      Obs        Mean    Std. Dev.       Min        Max
---------+-----------------------------------------------------------
  fecfat |        6    38.08333    22.47447        9.4       71.3

-> pilltype=   tablet
Variable |      Obs        Mean    Std. Dev.       Min        Max
---------+-----------------------------------------------------------
  fecfat |        6    16.53333    13.32091        4.6         38

-> pilltype= capsule
Variable |      Obs        Mean    Std. Dev.       Min        Max
---------+-----------------------------------------------------------
  fecfat |        6    17.41667    12.93745        3.4         36

-> pilltype=   coated
Variable |      Obs        Mean    Std. Dev.       Min        Max
---------+-----------------------------------------------------------
  fecfat |        6    31.06667     24.2641        5.8       68.2
```
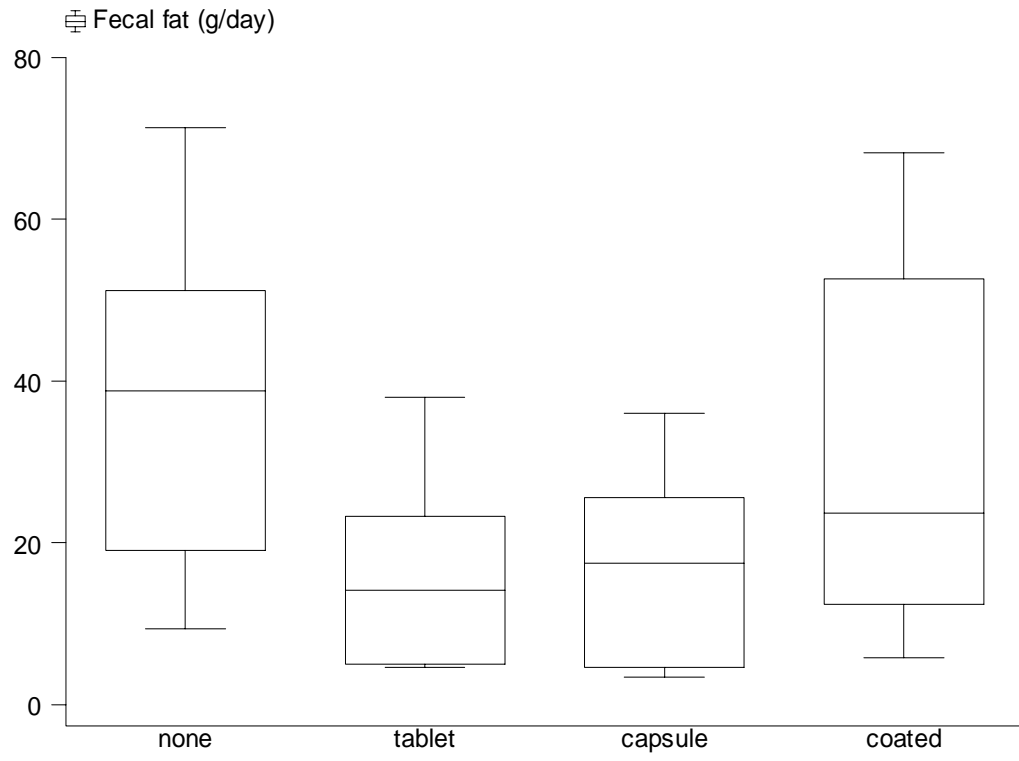
Analyses ignoring sex effects

ANOVA (wrong analysis)

```
. xi: regr fecfat i.pilltype
i.pilltype          Ipillt_1-4   (naturally coded; Ipillt_1 omitted)

  Source |       SS       df       MS              Number of obs =      24
---------+------------------------------          F(  3,    20) =    1.86
   Model |  2008.6017      3  669.533901           Prob > F      =  0.1687
Residual |  7193.36328    20  359.668164           R-squared     =  0.2183
---------+------------------------------          Adj R-squared =  0.1010
   Total |  9201.96498    23  400.085434           Root MSE      =  18.965


------------------------------------------------------------------------------
  fecfat |     Coef.   Std. Err.       t     P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
Ipillt_2 |    -21.55    10.9494     -1.968    0.063    -44.39005    1.29005
Ipillt_3 | -20.66667    10.9494     -1.887    0.074    -43.50672    2.173384
Ipillt_4 | -7.016668    10.9494     -0.641    0.529    -29.85672   15.82338
   _cons |  38.08333   7.742396      4.919    0.000     21.93298   54.23369
------------------------------------------------------------------------------

. testparm Ipill*

 ( 1)  Ipillt_2 = 0.0
 ( 2)  Ipillt_3 = 0.0
 ( 3)  Ipillt_4 = 0.0

       F(  3,    20) =    1.86
            Prob > F =    0.1687
```

Hierarchical analysis

```
. xi: xtgee fecfat i.pilltype, i(patid)
i.pilltype            Ipillt_1-4   (naturally coded; Ipillt_1 omitted)

Iteration 1: tolerance = 1.108e-15

GEE population-averaged model                Number of obs      =         24
Group variable:                       patid  Number of groups   =          6
Link:                              identity  Obs per group: min =          4
Family:                            Gaussian                 avg =        4.0
Correlation:                    exchangeable                 max =          4
                                             Wald chi2(3)       =      22.53
Scale parameter:                   299.7235  Prob > chi2        =     0.0001

------------------------------------------------------------------------------
  fecfat |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
Ipillt_2 |     -21.55    5.451781    -3.953   0.000    -32.23529   -10.86471
Ipillt_3 |  -20.66667    5.451781    -3.791   0.000    -31.35196   -9.981373
Ipillt_4 |  -7.016668    5.451781    -1.287   0.198    -17.70196    3.668626
   _cons |   38.08333    7.067808     5.388   0.000     24.23068    51.93598
------------------------------------------------------------------------------

. testparm Ipill*

 ( 1)  Ipillt_2 = 0.0
 ( 2)  Ipillt_3 = 0.0
 ( 3)  Ipillt_4 = 0.0

        chi2(  3) =    22.53
      Prob > chi2 =     0.0001
```

Hierarchical analysis (variation)

```
. xi: xtgee fecfat i.pilltype, i(patid) robust
i.pilltype              Ipillt_1-4   (naturally coded; Ipillt_1 omitted)

Iteration 1: tolerance = 1.662e-15

GEE population-averaged model              Number of obs      =         24
Group variable:                     patid  Number of groups   =          6
Link:                            identity  Obs per group: min =          4
Family:                          Gaussian                 avg =        4.0
Correlation:                  exchangeable                 max =          4
                                           Wald chi2(3)       =      11.71
Scale parameter:                  299.7235  Prob > chi2        =     0.0084

                        (standard errors adjusted for clustering on patid)
------------------------------------------------------------------------------
             |             Semi-robust
   fecfat |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
 Ipillt_2 |     -21.55   6.931847    -3.109   0.002    -35.13617   -7.96383
 Ipillt_3 |  -20.66667   7.349407    -2.812   0.005    -35.07124   -6.262094
 Ipillt_4 |  -7.016668   5.246295    -1.337   0.181    -17.29922   3.265881
    _cons |   38.08333   9.175163     4.151   0.000     20.10034   56.06632
------------------------------------------------------------------------------

. testparm Ipill*

 ( 1)  Ipillt_2 = 0.0
 ( 2)  Ipillt_3 = 0.0
 ( 3)  Ipillt_4 = 0.0

           chi2(  3) =    11.71
         Prob > chi2 =     0.0084
```

Analyses incorporating sex effects

ANOVA (wrong analysis)

```
. xi: regr fecfat i.pilltype i.sex
i.pilltype        _Ipilltype_1-4      (naturally coded; _Ipilltype_1 omitted)
i.sex             _Isex_0-1           (naturally coded; _Isex_0 omitted)

      Source |       SS       df       MS              Number of obs =      24
-------------+------------------------------           F(  4,    19) =    2.43
       Model |  3110.21668      4  777.554169           Prob > F      =  0.0837
    Residual |   6091.7483     19  320.618332           R-squared     =  0.3380
-------------+------------------------------           Adj R-squared =  0.1986
       Total |  9201.96498     23  400.085434           Root MSE      =  17.906


------------------------------------------------------------------------------
      fecfat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ipilltype_2 |     -21.55   10.33793    -2.08   0.051    -43.18753    .0875334
_Ipilltype_3 |  -20.66667   10.33793    -2.00   0.060     -42.3042    .970867
_Ipilltype_4 |  -7.016668   10.33793    -0.68   0.505     -28.6542   14.62087
     _Isex_1 |      13.55    7.31002     1.85   0.079    -1.750047   28.85005
       _cons |   31.30833   8.172851     3.83   0.001     14.20236   48.41431
------------------------------------------------------------------------------
```

Hierarchical analysis

```
. xi: xtgee fecfat i.pilltype i.sex, i(patid)
i.pilltype        _Ipilltype_1-4      (naturally coded; _Ipilltype_1 omitted)
i.sex             _Isex_0-1           (naturally coded; _Isex_0 omitted)

Iteration 1: tolerance = 1.219e-15

GEE population-averaged model             Number of obs      =        24
Group variable:                    patid  Number of groups   =         6
Link:                           identity  Obs per group: min =         4
Family:                         Gaussian                 avg =       4.0
Correlation:                 exchangeable                 max =         4
                                          Wald chi2(4)       =     24.00
Scale parameter:                 253.8228  Prob > chi2       =    0.0001


------------------------------------------------------------------------------
      fecfat |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ipilltype_2 |     -21.55   5.451781    -3.95   0.000    -32.23529   -10.86471
_Ipilltype_3 |  -20.66667   5.451781    -3.79   0.000    -31.35196   -9.981373
_Ipilltype_4 |  -7.016668   5.451781    -1.29   0.198    -17.70196    3.668626
     _Isex_1 |      13.55   11.16389     1.21   0.225    -8.330816    35.43082
       _cons |   31.30833   8.570992     3.65   0.000      14.5095    48.10717
------------------------------------------------------------------------------
```

Hierarchical analysis (variation)

```
. xi: xtgee fecfat i.pilltype i.sex, i(patid) robust
i.pilltype        _Ipilltype_1-4      (naturally coded; _Ipilltype_1 omitted)
i.sex             _Isex_0-1           (naturally coded; _Isex_0 omitted)

Iteration 1: tolerance = 1.219e-15

GEE population-averaged model              Number of obs      =         24
Group variable:                    patid   Number of groups   =          6
Link:                           identity   Obs per group: min =          4
Family:                         Gaussian                  avg =        4.0
Correlation:                 exchangeable                 max =          4
                                           Wald chi2(4)       =      12.80
Scale parameter:                253.8228   Prob > chi2        =     0.0123

                          (standard errors adjusted for clustering on patid)
-----------------------------------------------------------------------------
             |             Semi-robust
      fecfat |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+---------------------------------------------------------------
_Ipilltype_2 |     -21.55   6.931847    -3.11   0.002    -35.13617   -7.96383
_Ipilltype_3 |  -20.66667   7.349407    -2.81   0.005    -35.07124   -6.262094
_Ipilltype_4 |  -7.016668   5.246295    -1.34   0.181    -17.29922    3.265881
      _Isex_1 |      13.55   12.22942     1.11   0.268    -10.41923    37.51923
        _cons |   31.30833   4.918175     6.37   0.000     21.66889    40.94778
-----------------------------------------------------------------------------
```

Notes

- Hierarchical data structures are common.
- They lead to correlated data.
- Ignoring the correlation can be a serious error.

Fixed versus Random Factors

Definition: If a distribution is assumed for the levels of a factor it is <u>random</u>. If the values are fixed, unknown constants it is a <u>fixed</u> factor.

Ramifications:
- Scope of inference
  Inferences can be made on a statistical basis to the *population* from which the levels of the random factor have been selected.
- Incorporation of correlation in the model
  Observations that share the same level of the random effect are being modeled as correlated.
- Accuracy of estimates
  Using random factors involves making extra assumptions but gives more accurate estimates.
- Estimation method
  Different estimation methods must be used.

How to decide in practice?

SAS Proc MIXED philosophy:
fixed factors → MODEL statement
random factors → RANDOM statement
additional temporal and spatial correlation
→ REPEATED statement

SAS program for the Propranolol Example

```
data propran;
input bp patient upright drug;
cards;
96      1       0       0
71      1       0       1
73      1       1       0
87      1       1       1
96      2       0       0
85      2       0       1
104     2       1       0
76      2       1       1
            .
             .
              .
92      3       0       0
93      6       1       1
93      7       0       0
81      7       0       1
88      7       1       0
85      7       1       1

proc mixed;
     class patient upright drug;
     model bp=upright drug upright*drug;
     estimate "blup pat 1" | patient 1 ;
     estimate "blup pat 2" | patient 0 1;
     estimate "blup pat 3" | patient 0 0 1 ;
     estimate "blup pat 4" | patient 0 0 0 1;
     random patient;
     run;
```

SAS Output for the Propranolol Data

```
The SAS System          The MIXED Procedure

                Class Level Information
                Class       Levels  Values

                PATIENT        7   1 2 3 4 5 6 7
                UPRIGHT        2   0 1
                DRUG           2   0 1

          REML Estimation Iteration History
     Iteration  Evaluations     Objective     Criterion

            0            1   142.68756055
            1            1   141.94268164    0.00000000

                               Convergence criteria met.

         Covariance Parameter Estimates (REML)

             Cov Parm         Estimate

              PATIENT       15.79761905
              Residual      85.79761905

         Model Fitting Information for BP
          Description                        Value

       Observations                        28.0000
       Res Log Likelihood                 -93.0259
       Akaike's Information Criterion      -95.0259
       Schwarz's Bayesian Criterion        -96.2039
       -2 Res Log Likelihood              186.0517

       Tests of Fixed Effects

       Source            NDF    DDF   Type III F   Pr > F

       UPRIGHT             1     18        0.00     1.0000
       DRUG                1     18        7.70     0.0125
       UPRIGHT*DRUG        1     18        0.43     0.5221


          ESTIMATE Statement Results

Parameter        Estimate      Std Error     DF        t    Pr > |t|

blup pat 1      -3.86262200    3.17088923    18    -1.22     0.2389
blup pat 2      -0.25750813    3.17088923    18    -0.08     0.9362
blup pat 3      -0.99973746    3.17088923    18    -0.32     0.7562
blup pat 4       3.13554021    3.17088923    18     0.99     0.3358
```

Flowchart

Willing to assume the effects come from a distribution?

No -
FIXED   Yes -
                                          RANDOM

                        Interest lies in estimating

            Only the   Realized value
            distribution of   of random effect
            random effects.   and distribution.

Estimate variance   Estimate the variance
components.   components and
                                calculate predicted
                                values.
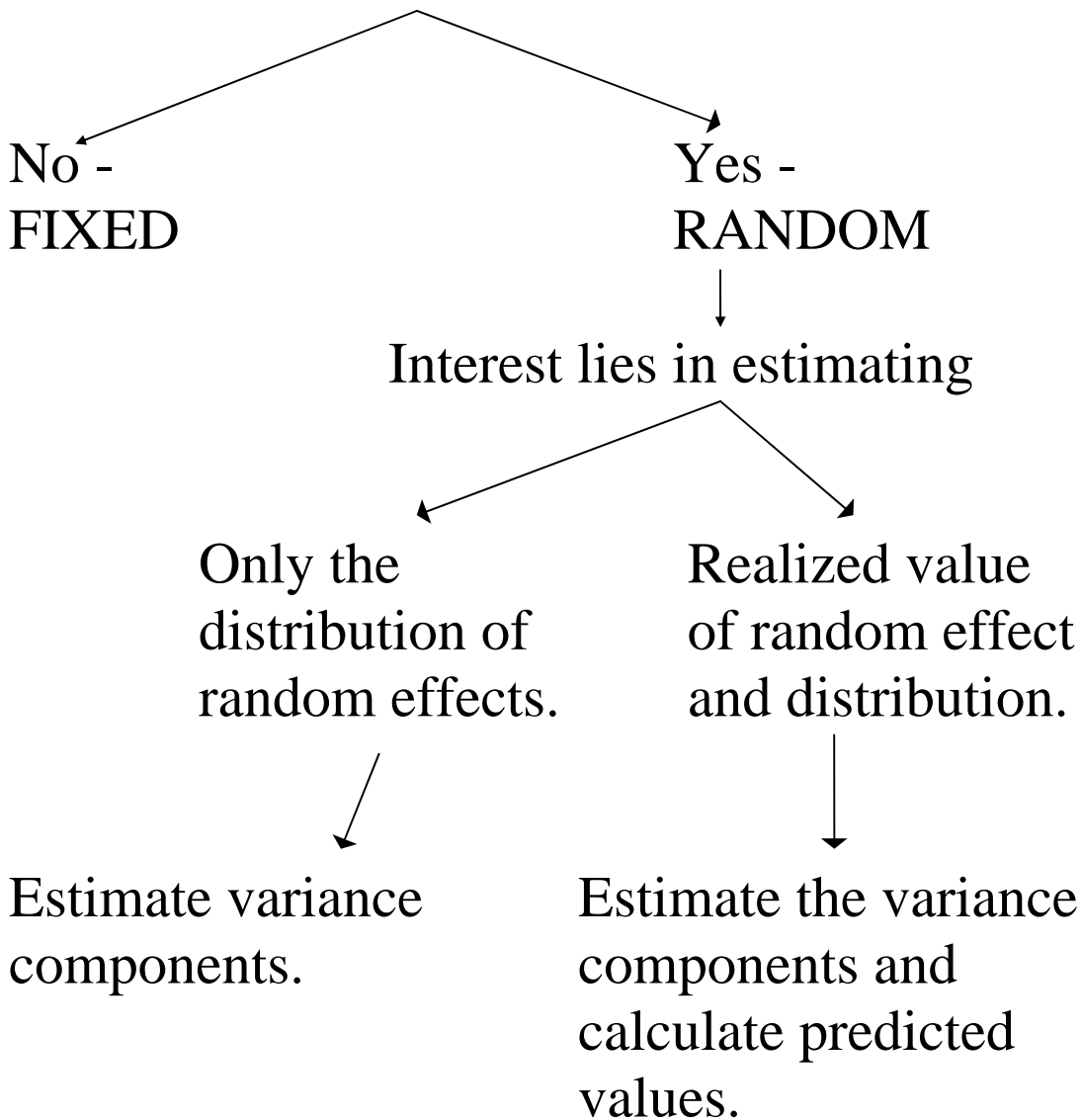
Assuming a factor is random involves extra assumptions but allows broader inferences.

Correlation in Mixed Models

Model:

Y_{ijk} = blood pressure for person k in condition (i,j).

$$= \mu + p_k + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

Covariance:

$$\text{Cov}(Y_{ijk}, Y_{i'j'k}) = \text{Cov}(p_k, p_k)$$
$$= \text{Var}(p_k)$$

$$\text{correlation} = \text{Var}(p_k)/[\text{Var}(p_k)+\text{Var}(\varepsilon_{ijk})]$$

Predicting the random effect

What if we assume a factor is random, but are interested in the individual levels of the random effects?

For the balanced data situation of the Propranolol data, the form of the best linear unbiased predictor is relatively simple and informative:

$$E[p_k | \mathbf{Y}] = E[p_k | \overline{Y}_{..k}] = \text{best predictor}$$

$$\begin{bmatrix} p_k \\ \overline{Y}_{..k} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ \overline{\mu} \end{bmatrix}, \begin{bmatrix} \sigma_p^2 & \sigma_p^2 \\ \sigma_p^2 & \sigma_p^2 + \sigma^2/n \end{bmatrix} \right),$$

where $\overline{\mu}$ is $\mu + \overline{\alpha}_. + \overline{\beta}_. + \overline{(\alpha\beta)}.$

$$E[p_k | \overline{Y}_{..k}] = E[p_k] + \text{cov}(p_k, \overline{Y}_{..k}) \text{var}(\overline{Y}_{..k})^{-1}(\overline{Y}_{..k} - \overline{\mu})$$

$$= 0 + \sigma_p^2(\sigma_p^2 + \sigma^2/n)^{-1}(\overline{Y}_{..k} - \overline{\mu})$$

$$= \frac{\sigma_p^2}{\sigma_p^2 + \sigma^2/n}(\overline{Y}_{..k} - \overline{\mu})$$

Best Linear Unbiased Prediction

$$\text{BLUP}(p_k) = \frac{\sigma_p^2}{\sigma_p^2 + \sigma^2/n}(\bar{Y}_{\cdot\cdot k} - \bar{Y}_{\cdot\cdot\cdot})$$

[Shrinkage]

Similarly,

$$\text{BLUP}(\bar{\mu} + p_k) = \bar{Y}_{\cdot\cdot\cdot} + \frac{\sigma_p^2}{\sigma_p^2 + \sigma^2/n}(\bar{Y}_{\cdot\cdot k} - \bar{Y}_{\cdot\cdot\cdot})$$

$$= \bar{Y}_{\cdot\cdot\cdot}\frac{\sigma^2/n}{\sigma_p^2 + \sigma^2/n} + \bar{Y}_{\cdot\cdot k}\frac{\sigma_p^2}{\sigma_p^2 + \sigma^2/n}$$

$$= \bar{Y}_{\cdot\cdot\cdot}\alpha + \bar{Y}_{\cdot\cdot k}(1\text{-}\alpha)$$

$$\alpha = \frac{\sigma^2/n}{\sigma_p^2 + \sigma^2/n}$$

[Weighted average]

In practice: EBLUP (Estimated BLUP)

$$\text{EBLUP}(p_k) = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}^2/n}(\bar{Y}_{\cdot\cdot k} - \bar{Y}_{\cdot\cdot\cdot})$$

Numerical illustration:

$$\text{EBLUP}(p_1) = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}^2/n}(\bar{Y}_{\cdot\cdot k} - \bar{Y}_{\cdot\cdot\cdot})$$

$$= \frac{15.7976}{15.7976 + 85.7976/4}(81.75 - 90.86)$$

$$= .424(-9.11)$$

$$= -3.863$$

$$\text{EBLUP}(\bar{\mu} + p_1) = \bar{Y}_{\cdot\cdot\cdot} + \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}^2/n}(\bar{Y}_{\cdot\cdot k} - \bar{Y}_{\cdot\cdot\cdot})$$

$$= 90.86 - 3.863$$

$$= 86.99$$

Contrast this with the mean for the first patient, which is 81.75.

BLUPs In Linear Mixed Models

The best predicted value of a random effect given the data is $\tilde{\mathbf{u}}$ = E[random effect|data].

A BLUP minimizes MSE of prediction among linear unbiased predictors:

$$\text{minimize } E[\,(\tilde{\mathbf{u}} - \mathbf{u})^2\,]$$

among $\tilde{\mathbf{u}}$ which are linear in Y and for which $E[\tilde{\mathbf{u}} - \mathbf{u}] = 0$.

For linear mixed models the best predictor is

$$\tilde{\mathbf{u}}_{BP} = \mathbf{DZ'V^{-1}(Y\text{-}X\beta)},$$

while the best linear unbiased predictor is

$$\tilde{\mathbf{u}}_{BLUP} = \mathbf{DZ'V^{-1}(Y-X\hat{\beta})}$$

"Shrinkage" estimator.

<u>Bottom line</u>: can be interested in the specific levels of a random factor.

Estimation and Tests in LMMs

Estimation of parameters by maximum likelihood or restricted maximum likelihood. Maximize the log of the likelihood.

Tests of fixed effects via approximate F- tests (SAS PROC MIXED).

Basic idea: Consider H_0: $\mathbf{k}'\boldsymbol{\beta} = 0$.

Could do a likelihood ratio test or a Wald test.

$$\text{var}(\mathbf{k}'\hat{\boldsymbol{\beta}}) \cong \mathbf{k}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\mathbf{k}$$

$$\frac{\mathbf{k}'\hat{\boldsymbol{\beta}}}{\sqrt{\mathbf{k}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\mathbf{k}}} \sim N(0,1) \text{ under } H_0$$

But need $\hat{\mathbf{V}}$ in place of $\mathbf{V}$.

Distribution?

Tests of variances of random effects

When using a maximum likelihood analysis the typical tests are based on the improvement in the maximized value of the log likelihood. The difference in twice the log likelihood is compared to a chi-square distribution to test for statistical significance. For testing whether a variance component is equal to zero the usual method must be slightly modified. Ordinarily we would take the difference in log likelihoods of the models with and without the random effect and compare that directly to a χ^2_1 cutoff point. The modification is to either calculate a p-value and then cut it in half, or to compare to a cutoff point with twice the nominal α level.

Why? The intuition is that testing

$$H_0: \ \sigma^2_p = 0 \text{ versus } H_0: \ \sigma^2_p > 0$$

is a one-sided test. The usual test is inherently two-sided and must be adjusted to reflect this fact.

3. Review: Generalized Linear Models (GLMs)

Example: (from Finney, *Statistical Method in Bioassay, 3rd Ed.*). Study of growth of *Bacillus mesentericus* spores grown in dilutions of a potato flour suspension.

Dilution (g/100ml)	Spore Growth Number of plates	Number positive	Proportion
1/128	5	0	0.0
1/64	5	0	0.0
1/32	5	2	0.4
1/16	5	2	0.4
1/8	5	3	0.6
1/4	5	4	0.8
1/2	5	5	1.0
1	5	5	1.0
2	5	5	1.0
4	5	5	1.0

Analysis?

Plot of Potato Flour Data
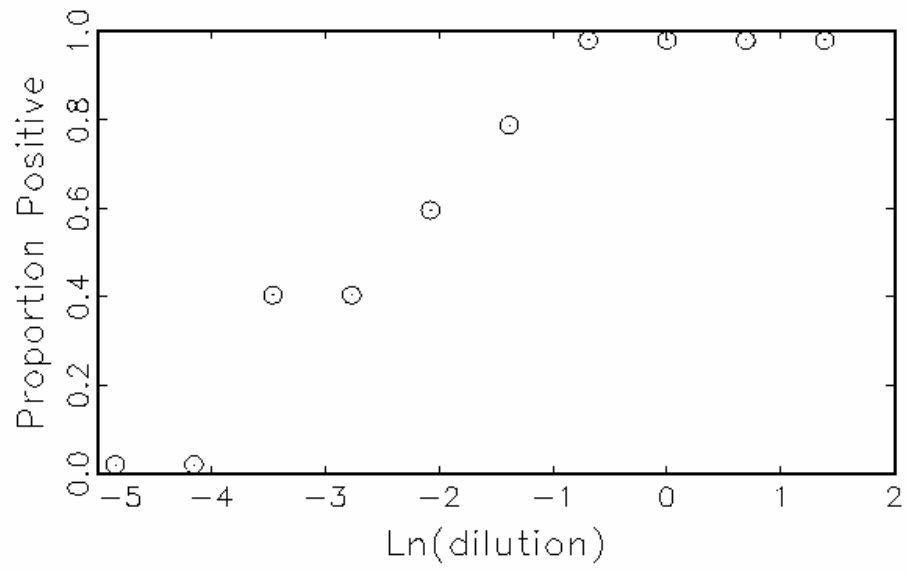
Analysis of potato flour data

Logistic Regression

Notation: Let x_i be the ln(dilution) for the i*th* series and let Y_i be the number of positive plates.

Distribution: $Y_i \sim$ indep. Binomial(5, $p(x_i)$), which has mean $5p(x_i)$.

Model: $\ln(p(x_i)/(1-p(x_i))) = \alpha + \beta x_i$

S-shaped function of x

When $x = -\alpha/\beta$, $\alpha + \beta x = 0$, and $p(x) = 1/2$.

<u>Loglikelihood</u>: $\Sigma_i y_i(\alpha+\beta x_i)-\log(1+\exp(\alpha+\beta x_i))$

<u>Maximum likelihood estimates</u>:

$\hat{\alpha} = 4.17$

$\hat{\beta} = 1.62$

for ln(dilution) which achieves 50%

positive results: -4.17/1.62 = -2.57

$\exp(-2.57) = 0.076 \approx 1/13$
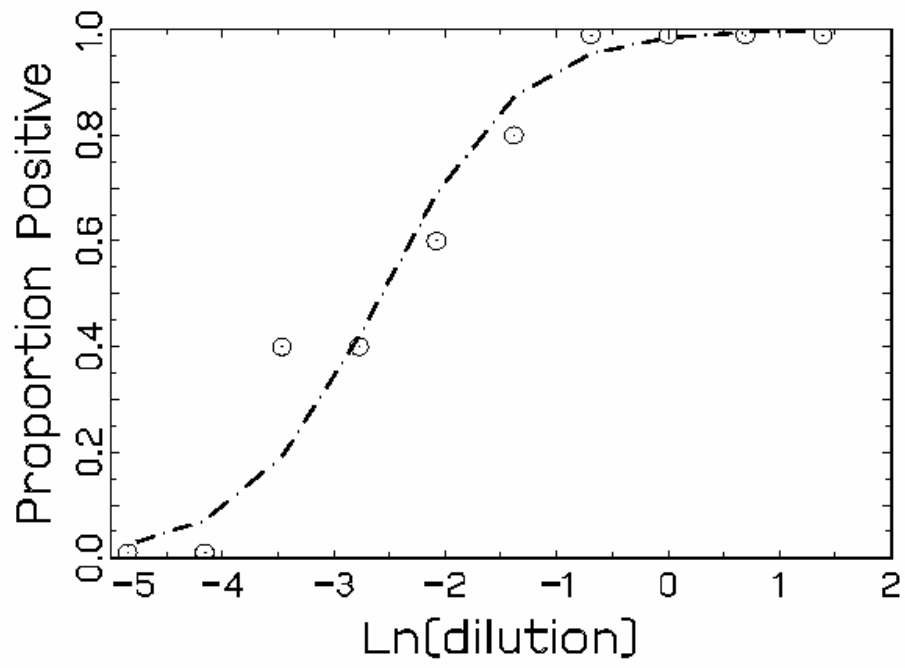
GLMs

Dissect the modeling process into three distinct components:

    1. What is the distribution of the data?

    2. What aspect of the problem will be
        modelled?

    3. What are the predictors?

<u>In our example:</u>

    1. No. of successes in 5 trials => Binomial

    2. log odds = $\ln(p/(1-p))$

    3. ln(dilution)

GLMs

general case	our example
Y ~ distribution	Y ~ Binomial
μ = mean of Y	np = mean of Y
$g(\mu) = X\beta$	$\ln(p/(1-p)) = \alpha + \beta x$
link function $g(\cdot)$	logit link
covariates $X\beta$	one predictor x

Example: (from Snedecor and Cochran, Sec 16.9, via McCullagh and Nelder, *Generalized Linear Models*). Amount of ascorbic acid remaining in snap beans after 2,4,6, and 8 weeks of storage at 0, 10 or 20 ^{o}F (a 3×4 factorial with 3 replicates per treatment combination).

Sum of three ascorbic acid determinations for each of 12 treatments on snap beans

Temp	Weeks of storage				Average
	2	4	6	8	
0	45	47	46	46	46.0
10	45	43	41	37	41.5
20	34	28	21	16	24.8
Ave	41.3	39.3	36.0	33.0	37.4

Here is a graph of the results.

Plot of Total Ascorbic Acid Versus Week
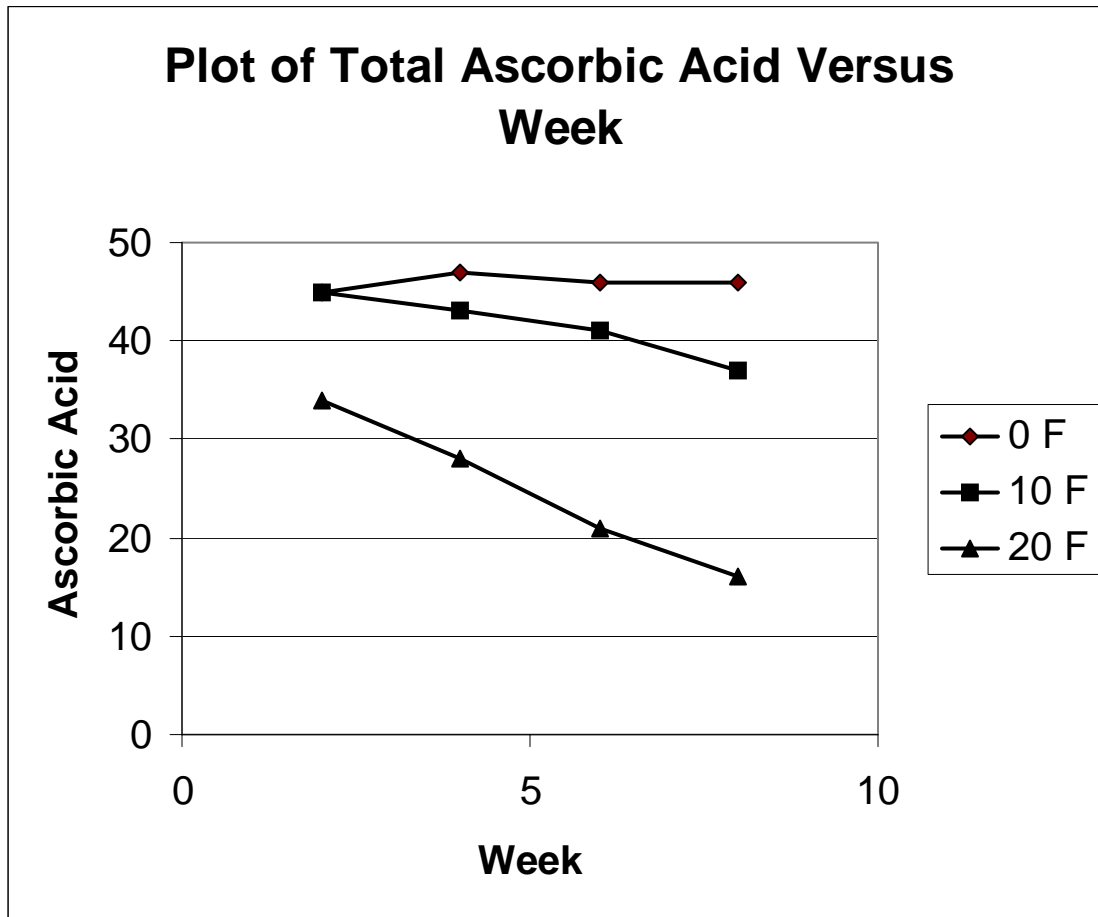
<u>Analysis</u>: McCullagh and Nelder assume that the variance in ascorbic acid determination is constant on the original scale and wish to fit a model with exponential decline through time.

If Y_{ij} = average value at the i*th* temperature for week t_j, then a possible model is

$$Y_{ij} \sim \text{Normal}(\exp\{\alpha - \beta_i t_j\}, \sigma^2)$$

This is a generalized linear model for a Normal distribution with constant variation and with log link. The model has a common intercept and different slopes through time for each storage temperature.

Another possibility: log transform.

$$\ln(Y_{ij}) \sim \text{Normal}(\alpha - \beta_i t_j, \tau^2)$$

Transform or Link?

Example: Average daily fat yield (kg/day) from milk from a single cow for each of 35 weeks.

0.31	0.39	0.50	0.58	0.59	0.64
0.68	0.66	0.67	0.70	0.72	0.68
0.65	0.64	0.57	0.48	0.46	0.45
0.31	0.33	0.36	0.30	0.26	0.34
0.29	0.31	0.29	0.20	0.15	0.18
0.11	0.07	0.06	0.01	0.01	

Plot of Fat Yield (kg/day) for Each Week
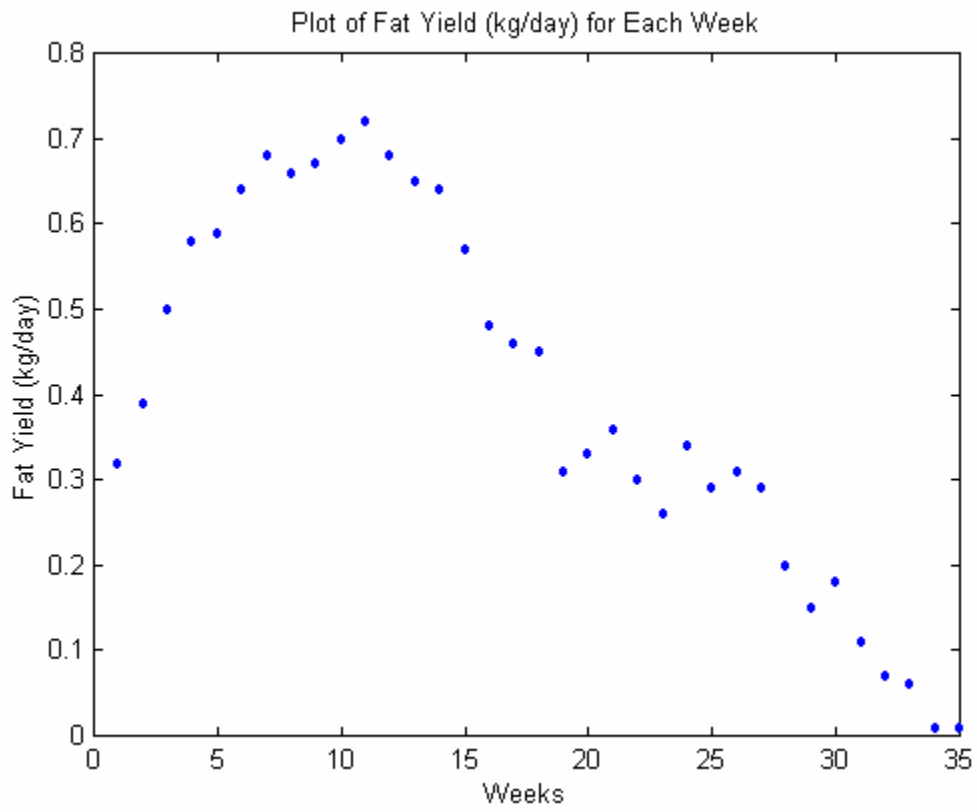
A typical model:

$$\text{Fat yield "=" } \alpha t^{\beta} e^{\gamma t} \quad \text{where t=week}$$

Transform:

$$\ln Y_i \sim N(\ln(\alpha) + \beta\ln(t_i) + \gamma t_i, \sigma^2)$$
$$\ln Y_i = \ln(\alpha) + \beta\ln(t_i) + \gamma t_i + \varepsilon_i$$
$$E[\ln Y_i] = \ln(\alpha) + \beta\ln(t_i) + \gamma t_i$$
$$Y_i = \alpha t_i^{\beta} e^{\gamma t_i} e^{\varepsilon_i}$$

Link:

$$Y_i \sim N(\alpha t_i^{\beta} e^{\gamma t_i}, \tau^2)$$
$$E[Y_i] = \alpha t_i^{\beta} e^{\gamma t_i}$$
$$\ln(E[Y_i]) = \ln(\alpha) + \beta\ln(t_i) + \gamma t_i$$
$$Y_i = \alpha t_i^{\beta} e^{\gamma t_i} + \delta_i \qquad \delta_i \sim N(0, \tau^2)$$

Plot of Fat Yield (kg/day) for Each Week

Plot of Ln Fat Yield (kg/day) for Each Week
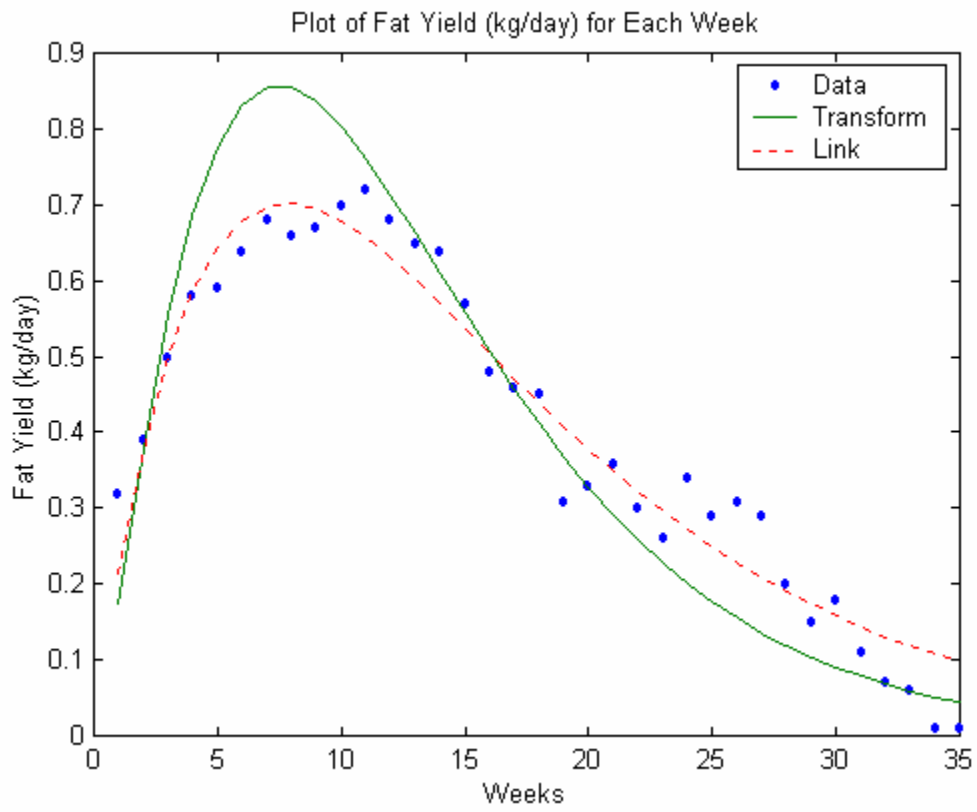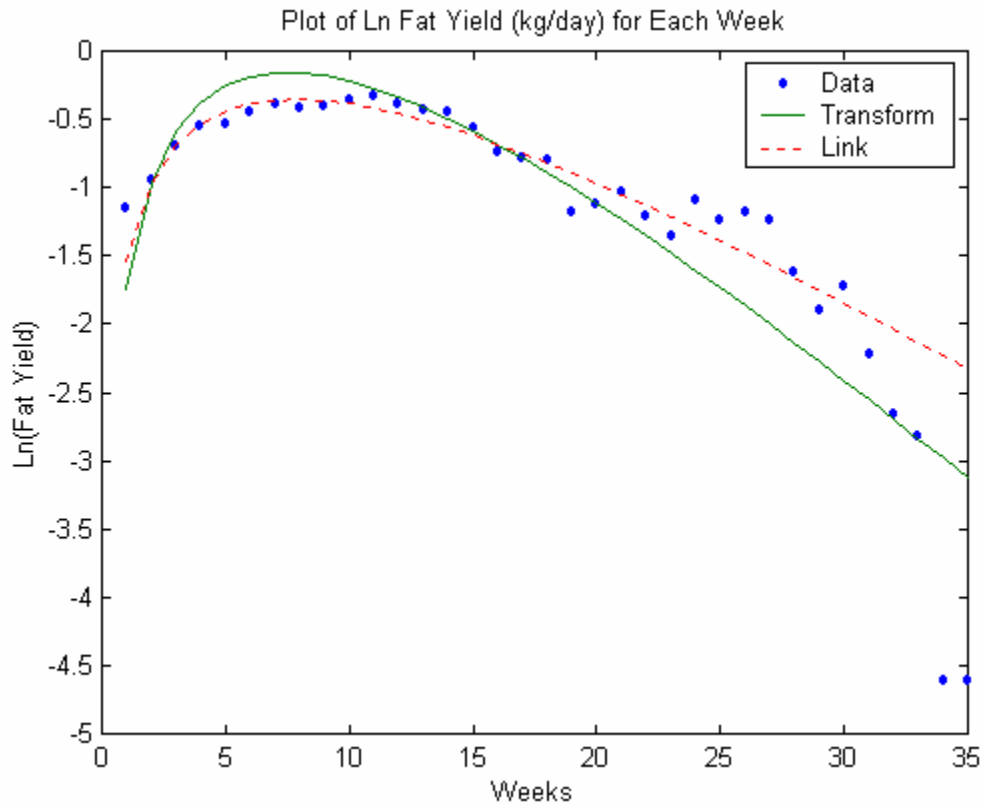
Homoscedasticity: The GLM analysis assumes a constant variance on the original scale. The transformed analysis assumes a constant variance on the transformed scale.

Trouble with transformations: With Poisson distributed data with zero counts, using a link function avoids the problems of a log transformation and zero counts.

See Ruppert, Cressie, and Carroll (1989) for a discussion.

Estimation and Tests in GLMs

Estimation of parameters by maximum likelihood or maximum quasi-likelihood. Maximize the log of the likelihood or quasi-likelihood.

Quasi-likelihood estimation:

Suppose $Var(Y) = \sigma^2 V(\mu)$

Define $U = \dfrac{Y - \mu}{\sigma^2 V(\mu)}$

and $Q(\mu;y) = \displaystyle\int_y^\mu \frac{y - t}{\sigma^2 V(t)} dt$

Note that
$$E[U] = 0$$
$$Var(U) = \frac{1}{\sigma^2 V(\mu)}$$
$$-E[\frac{\partial U}{\partial \mu}] = \frac{1}{\sigma^2 V(\mu)}$$

which is similar to the properties of $\dfrac{\partial \ln f_{Y_i}}{\partial \mu}$.

For maximum likelihood we solve

$$\frac{\partial \ln L}{\partial \mu} = \sum_i \frac{\partial \ln f_{Y_i}}{\partial \mu} = 0.$$

For maximum quasi-likelihood we solve

$$\frac{\partial}{\partial \mu} \sum_i Q(\mu_i, y_i) = 0.$$

<u>Example:</u>

$$\sigma^2 = 1, \ V(\mu) = \mu$$

$$U = \frac{Y - \mu}{\mu}$$

and $Q(\mu; y) = \int_y^\mu \frac{y - t}{t} dt$

$$= y \int_y^\mu \frac{1}{t} dt - \int_y^\mu dt$$

$$= y \ln(\mu) - y \ln(y) - (\mu - y)$$

So $\sum_i Q(\mu, y_i) = \sum_i y_i \ln(\mu) - n\mu + \text{constant}$.

For a Poisson,

$$\ln L = \sum_i Q(\mu, y_i) = \sum_i y_i \ln(\mu) - n\mu + \text{constant}$$

Measure of model (lack of) fit: deviance or Pearson chi-square statistic.

Deviance = 2(max possible loglikelihood - loglikelihood of fitted model)

So large values of deviance indicate a model which fits poorly.

Difference in Deviance for models 1 and 2 = 2(loglik model 2 - loglik model 1) = likelihood ratio statistic

<u>Example</u>: Potato flour dilutions (continued)

Maximum achievable loglikelihood = -12.597

<u>Model 1</u>: $logit(p(x_i)) = \alpha + \beta x_i$
    ML estimates: $\hat{\alpha} = 4.1737$
                    $\hat{\beta} = 1.6226$
    loglikelihood: -14.214

    Deviance: 2(-12.597+14.214)
                    = 3.234 with 10-2 = 8 d.f.

<u>Model 2</u>: $logit(p(x_i)) = \alpha$ (no slope)
    ML estimate: $\hat{\alpha} = 0.4896$
       Note: 1/(1+exp(-0.4896))=.62=ave prop.

    loglikelihood: -33.203

    Deviance: 2(-12.597+33.203)
                    = 41.212 with 10-1 = 9 d.f

Difference in deviance = 41.212 - 3.234
                      = 37.978 with 1 d.f.

Software

Software for LMMs and GLMs is readily available, either through special purpose routines, e.g., for logistic regression, or general routines. The package GLIM was the pioneer of software for GLMs, but other packages, e.g., SAS have caught up and now offer GLMs.

In SAS, PROC MIXED fits linear mixed models with the assumption of normality and PROC GENMOD fits generalized linear models.

Analysis of the potato flour data using GENMOD:

Program:

```
data one;
set work.potflour;
lndil=log(dilution);
run;
proc genmod descending;
model nopos/noplate=lndil/dist=bin;
run;
```

Output:

```
                          The GENMOD Procedure

                           Model Information

        Data Set                       WORK.ONE
        Distribution                   Binomial
        Link Function                     Logit
        Response Variable (Events)        Nopos     Nopos
        Response Variable (Trials)      Noplate   Noplate
        Observations Used                    10
        Number Of Events                     31
        Number Of Trials                     50


                  Criteria For Assessing Goodness Of Fit

            Criterion              DF         Value      Value/DF

            Deviance               8        3.2329        0.4041
            Scaled Deviance        8        3.2329        0.4041
            Pearson Chi-Square     8        2.7175        0.3397
            Scaled Pearson X2      8        2.7175        0.3397
            Log Likelihood                -14.2136


        Algorithm converged.


                       Analysis Of Parameter Estimates

                               Standard   Wald 95% Confidence      Chi-
    Parameter    DF   Estimate     Error        Limits           Square   Pr > ChiSq

    Intercept     1     4.1737    1.2522    1.7194    6.6280      11.11      0.0009
    lndil         1     1.6226    0.4571    0.7266    2.5185      12.60      0.0004
    Scale         0     1.0000    0.0000    1.0000    1.0000

    NOTE: The scale parameter was held fixed.
```

4. Introduction to Generalized Linear Mixed Models (GLMMs)

Example: Similar to Abu-Libdeh, Turnbull, Clark, (Biometrics, 1990).   Effect of selenium on prevention of skin cancer. 770 patients from seven clinics followed for four years.

Recorded:

    response: Number of new basal cell epithelioma (BCE) sites found.

    predictors: Selenium? (SEL), Sex (SEX), Exposure to the sun (SUN).

    [Also?]   Age, childhood farm exposure?, smoker?, skin damage, no. of tumors previously, clinic...

        Q1: Does selenium decrease the number of BCEs?

        Q2: Are some patients more sensitive to sun exposure? If so, which ones?

Features from a modelling viewpoint

Nature of response: count data

How to relate the response to the predictors?

λ=mean

$\ln(\lambda) = \mu + \beta_1 SEX + \beta_2 SEL + \gamma SUN$

=> Poisson regression

Problems: 1.

2.

A Generalized Linear Mixed Model

Let Y_{ij} be the response for patient i at visit j.

$$Y_{ij} = \text{Number of new BCE sites}$$

Assume $Y_{ij} \sim \text{Poisson}(\lambda_{ij})$, where λ_{ij} is the mean number of new BCEs for patient i at visit j.

$$\ln(\lambda_{ij}) = \mu_i + \beta_1 \text{SEX} + \beta_2 \text{SEL} + \gamma \text{SUN}$$

$$\mu_i \sim \text{Normal}(\mu, \tau_\mu)$$

assume a distribution for μ_i

$$\text{Cov}(\ln(\lambda_{ij})), \ln(\lambda_{ik})) = \tau_\mu$$

A correlation is induced in the model between observations taken on the same patient.

Other features

1. <u>Assume a distribution on γ:</u>

$\ln(\lambda_{ij}) = \mu_i + \beta_1 SEX + \beta_2 SEL + \gamma_i SUN$

From the previous model:

γ = sun exposure effect (same across patients)

γ_i = sun exposure effect for the i*th* patient

$\gamma_i \sim Normal(\gamma, \tau_\gamma)$

- $\tau_\gamma > 0$ <=> patients have different responses.
- Extreme values of γ_i indicate sensitive individuals.

2. <u>Assume a distribution on β_2:</u>

$\ln(\lambda_{ij}) = \mu_i + \beta_1 SEX + \beta_{2i} SEL + \gamma SUN$

$\beta_{2i} \sim Normal(\beta_2, \tau_\beta)$

- If SEL is coded 1 for yes and 0 for no, then for the placebo group, the contribution of the $\beta_{2i}SEL$ term is zero, while for the treatment group it is β_{2i}. If $\tau_\beta > 0$, then the treatment group will have a larger variance.

Specifying GLMMs

1. What is the distribution of the data?

2. What aspects will be modelled?

3. What are the factors?

*   4. Which factors will be assumed to have a

        distribution?

GLMMs

general case	logit-normal
Y ~ distribution	Y ~ Bernoulli
μ = mean of Y	p = mean of Y
$g(\mu) = X\beta + Zu$	$\ln(p/(1-p)) = \beta x + u_i$
link function $g(\cdot)$	logit link
fixed factors $X\beta$	fixed factor x
random factors Zu	random intercepts u_i
u ~ distribution	$u_i \sim Normal(\mu_u, \tau_u)$

Prediction in GLMMs

In GLMMs we can adopt the same strategy as in LMMs:

(1) Calculate $\tilde{\mathbf{u}} = E[\,\mathbf{u}\,|Y]$

(2) Estimate any unknown parameters

However, either of these steps may be problematic.

5. Modeling in GLMMs

Example 1: Progabide and seizures (Diggle, Liang and Zeger, 1994).

Epileptics were randomly allocated to a placebo or an anti-seizure drug (Progabide) group. The number of seizures was recorded for a baseline period of 8 weeks and during consecutive two-week periods for four periods after beginning treatment. Is the drug effective at reducing the number of seizures?

<u>Example 2</u>: Cartoons and learning disabilities

This study concerned the comprehension of humor in two groups of adolescents (normal and learning disabled). Each subject was exposed to 24 different cartoons (in three types). There are two response variables whether or not the child got the cartoon and whether or not s/he liked it. The types of cartoon are: visual only, linguistic only, and both visual and linguistic.

Two questions of interest are: Is there a difference between normal and learning disabled children? How consistent are the responses within cartoon type?

<u>Example 3</u>: Photosynthesis in corn relatives

Two species of corn relatives (an annual and perennial) are being compared with respect to photosynthetic physiology. Seeds from two populations of each species were collected and grown in the greenhouse. The experimental design was a randomized complete block design with four blocks and three seeds from each population in each block (for a total of 12 seeds per block) After 24 days, photosynthesis was recorded at nine different light levels from full sunlight to darkness on one individual from each population in each block (N=16). Measurements on the same 16 plants were repeated after 48 days. From these data, photosynthesis versus irradiance (PAR) response curves reflecting the change in photosynthetic rate with light level were derived.

The traits of interest are the maximum photosynthetic rate, dark respiration, the light compensation point, and the quantum yield. The maximum photosynthetic rate measures the maximum amount of carbon dioxide the plants

are able to assimilate in full sunlight, the dark respiration indicates how much carbon dioxide they respire in the dark, the light compensation point is the light level at which photosynthesis overcomes respiration and carbon assimilation becomes positive, and quantum yield is the efficiency of carbon assimilation at low light levels, or the slope of the light response curve as it crosses the light compensation point.

The main question of interest is to compare the two species with respect to their photosynthetic traits

Photosynthetic Rate Versus Light for Two Plants
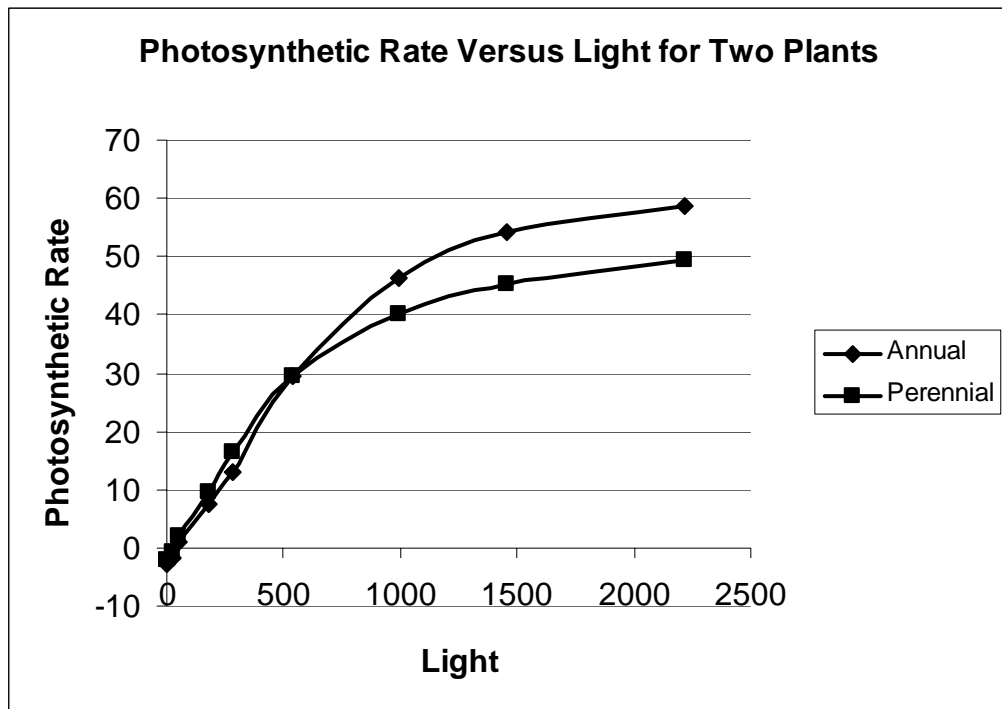
<u>Example 4</u>: Chestnut Leaf Blight

The American chestnut tree was a predominant hardwood in the forests of the eastern United States, reaching 80-100 feet in height at maturity and providing timber and low-fat, high-protein nutrition for animals and humans in the form of chestnuts. In the early 1900's an imported fungal pathogen, which causes chestnut leaf blight, was introduced into the United States. The pathogen spread from infected trees in the New York City area and, by 1950, had killed over 3 billion trees and virtually eliminated the chestnut tree in the United States. Economic losses in both timber and nut production have been estimated in the hundreds of billions of dollars. As well, there are ecological impacts of eliminating a dominant species.

Attempts to restore this tree to the U.S. forests include
• development of blight resistant varieties
• weakening of the fungus by infecting it with a virus which reduces the fungus' virulence.

I'll describe the latter in more detail. The basic idea is to release hypovirulent isolates of chestnut blight fungus and let the viruses infect the natural populations of the fungus, thereby allowing chestnuts trees to survive.

Viruses spread between fungal individuals when they come in contact and fuse together. A major obstacle in spreading this virus and thus controlling the disease is that different isolates of the fungus cannot necessarily transfer the virus to one another.

Michael Milgroom - Cornell Plant Pathology, and his colleague, Paolo Cortesi - from the University of Milan, have worked with six incompatibility genes, which may block the transmission of this virus between isolates of the fungus.

To estimate the effects of these genes, they have paired numerous isolates which differ on the first gene only, the second gene only, the first and the second gene, etc. For each combination of isolates they have averaged about 30 attempts and record a binary response of whether or not the attempt succeeded in transmitting the virus.

<u>Questions of interest</u> include whether pre-identified genes actually do have an influence on transmission of the virus (and if so, to what degree), whether there are other, as yet unidentified, genes which might affect transmission, and whether transmission is symmetric. By symmetry of transmission we mean the following: suppose the infected fungus is type b at the locus for the first gene and the non-infected isolate (which we are trying to infect) is type B. The two isolates are the same at the other five loci. Is the probability of transmission the same as when using a type B to try to infect a type b?

<u>Example 5</u>: Combat vehicle design

Army combat vehicles of the future will likely locate crew stations deep within the vehicle, to achieve lower silhouettes and increased crew protection against ballistic and directed energy threats. This will require indirect vision systems such as liquid crystal displays.

In "Indirect Vision Driving With Fixed Flat Panel Displays for Near-Unity, Wide, and Extended Fields of Camera View" (Smyth, Gombash, Burcham, ARL-TR-2511, 2001) eight drivers tested each of four vision systems: direct and three types of indirect (with three different fields of view -- unity, wide and extended). Outcomes included speed to traverse the course, number of barrels knocked over and severe motion sickness (yes/no).

<u>Example 6</u>: Troponin and hemorrhage

Heart damage in patients experiencing brain hemorrhage has historically been attributed to pre-existing conditions. However, more recent evidence suggests that the hemorrhage itself can cause heart damage through the release of norepinephrine following the hemorrhage. To study this, researchers at UCSF measured cardiac troponin levels, an enzyme released following heart damage, at up to three occasions after patients were admitted to the hospital for a specific type of brain hemorrhage (subarachnoid hemorrhage or SAH).

The primary question was whether severity of injury from the hemorrhage was a predictor of troponin levels, as this would support the hypothesis that the SAH caused the cardiac injury. To make a more convincing argument in this observational study, we would like to show that severity of injury is an independent predictor, over and above other circulatory and clinical factors that would predispose the patient to higher troponin levels.

Possible clinical predictors included age, gender, history of heart failure, heart rate, whether the person was a smoker, diabetic or had high cholesterol levels. Circulatory status was described using systolic blood pressure, history of hypertension (yes/no) and left ventricular ejection fraction a measure of heart function. The severity of neurological injury was graded using a subject's Hunt-Hess score on admission. This score is an ordered categorical variable ranging from 1 (little or no symptoms) to 5 (severe symptoms such as deep coma).

The study involved 175 subjects with at least one troponin measurement and between 1 and 3 visits per subject.

6. Features of GLMMs

6 a) Consequences of model assumptions

What impact does this have on the distribution of Y? Here are some calculations for the skin cancer example.

$$E[Y_{ij}] = E[E[Y_{ij}| \mu_i]]$$
$$= E[\exp\{\mu_i + \beta_1 SEX + \beta_2 SEL + \gamma SUN\}]$$
$$= \exp\{\beta_1 SEX + \beta_2 SEL + \gamma SUN\}E[\exp\{\mu_i\}]$$

So $\log E[Y_{ij}] = \beta_1 SEX + \beta_2 SEL + \gamma SUN + \log M_\mu(1)$, where $M_\mu(t)$ is the moment generating function of μ_i.

$$Var(Y_{ij}) = Var[E[Y_{ij}| \mu_i]] + E[Var(Y_{ij}| \mu_i)]$$
$$= Var(E[\exp\{\mu_i + \beta_1 SEX + \beta_2 SEL + \gamma SUN\}])$$
$$+ \exp\{\beta_1 SEX + \beta_2 SEL + \gamma SUN\}E[\exp\{\mu_i\}]$$
$$= Var(E[\exp\{\mu_i + \beta_1 SEX + \beta_2 SEL + \gamma SUN\}])$$
$$+ E[E[Y_{ij}| \mu_i]]$$
$$> E[Y_{ij}]$$

Marginal distribution for Probit models

$Y_{ij} \sim$ Bernoulli($\Phi[$ $\mu + a_i + \beta x_{ij}]$)

$a_i \sim$ Normal($0, \tau_a$).

What is the marginal distribution?

Persistence of links

Which other links "persist" like this?

<u>Log link</u>:
$$E[Y_i] = E[E[Y_i|\ \mathbf{u}\ ]]$$
$$= E[\exp\{\ \mathbf{x}_i'\boldsymbol{\beta}+\mathbf{z}_i'\mathbf{u}\ \}]$$
$$= \exp\{\ \mathbf{x}_i'\boldsymbol{\beta}\ \}E[\exp\{\ \mathbf{z}_i'\mathbf{u}\ \}]$$

So $\log E[Y_i] = \mathbf{x}_i'\boldsymbol{\beta} + \log E[\exp\{\ \mathbf{z}_i'\mathbf{u}\ \}]$
(partial)

Logit link and other links do not persist.

6 b) Marginal versus conditional models

Illustration of difference of conditional and marginal approaches:

$Y_{ij} = 1$ if the i*th* woman miscarries during her j*th* pregnancy and is 0 otherwise.
$x_{ij} = j$ = pregnancy number.

Model:

$$E[Y_{ij}|u_i] = \Phi(\mu+\beta x_{ij}+u_i)$$

which gives

$$E[Y_{ij}] = \Phi\left(\frac{\mu+\beta x_{ij}}{\sqrt{1+\sigma^2_u}}\right)$$

Interpretation of β?

Advantages/disadvantages of the approaches

Because of computational problems with conditionally specified GLMMs there are many alternative methods (e.g., GEEs) for clustered data that focus on models for the marginal expectation of the response, $E[\ Y_{ij}\ ]$.

Marginal models have the following advantages:
- Marginal models avoid the specification of the conditional structure, so misspecification of this portion of the model can be avoided.
- For example, when the underlying random effects distribution is heteroscedastic, assuming it is homoscedastic and using a conditional approach can lead to biased estimators (Heagerty and Kurland, 2001)
- When paired with a GEE approach to estimation, estimates of the marginal parameters are consistent, even under misspecification of the association structure.

Major drawbacks of the marginal approach include:

- Often does not measure covariate effects of primary scientific interest.
- In extreme circumstances, features of scientific interest present in every conditional model may not be present in the marginal model.
- Marginal quantities can be calculated from a conditional model but the converse is not typically true.
- Marginal modeling approaches are susceptible to Simpson's paradox and the Ecological Fallacy, potentially giving misleading results.
- If the question of interest is based on the marginal distribution, a longitudinal design may not be the most appropriate.

For a more detailed critique of marginal modeling see Lindsey and Lambert (1998).

7. Inference for GLMMs

<u>Estimation</u>: Maximum likelihood (or variants) based on normality assumptions are relatively standard for linear mixed models. For example, SAS PROC MIXED using ML or REML.

For many GLMs, maximum likelihood is also standard, e.g., logistic regression or Poisson regression.

What about GLMMs?

A simple GLMM

<u>A logit-normal model</u>:

$$Y_{ij} \mid u \sim \text{Bernoulli}(p_{ij}),$$
$$i=1,2, ...n; \ j=1,2, ...q.$$
q clusters, n observations per cluster.

$$\ln(p_{ij}/(1-p_{ij})) = \beta x_{ij} + u_j$$
logit link
one fixed and one random factor

$$u_j \sim \text{Normal}(0,\sigma^2)$$

<u>Scenario</u>:
$Y_{ij} = 1$ if blood pressure on day i on individual j decreases after using medicine at dose x_{ij}, 0 otherwise.

q individuals, n days of measurement on each.

u_j is the individual specific propensity to increase or decrease blood pressure.

ML Estimation?

$$\text{Likelihood} = P\{Y=y|\beta, \sigma^2\}$$

$$= \int P\{Y=y|\beta, \sigma^2, u\}f(u)du$$

$$= \int P\{Y=y|\beta, u\}f(u)du$$

$$= \int \prod_{i,j} P\{Y_{ij} = y_{ij}|\beta, u\} \; f(u)du$$

$$= \prod_j \int \prod_i P\{Y_{ij} = y_{ij}|\beta, u_j\}f(u_j)du_j$$

$$=$$

$$\prod_j \int \exp\{\beta\Sigma_i Y_{ij}x_{ij} + Y_{+j}u_j\}\prod_i(1+\exp\{\beta x_{ij}+u_j\})^{-1}\times$$

$$\exp\{-u_j^2/2\sigma^2\}/(2\pi\sigma^2)^{1/2}du_j.$$

Cannot be evaluated in closed form but is not too hard to do numerically for this example.

Brute force ML

When the model has a single random effect or two nested random effects, it is relatively easy to evaluate the integrals in the likelihood. For example, with a single random factor we have seen that the likelihood is a product of one-dimensional integrals.

One can then maximize the likelihood numerically to find ML estimates and to perform likelihood ratio tests.

Numerical evaluation of the likelihood

When there is a single, normally distributed random effect, the likelihood can be written as a product of integrals of the form:

$$\int_{-\infty}^{+\infty} g(x) \exp\{-x^2\} dx$$

These can be accurately evaluated using Gauss-Hermite quadrature:

$$\int_{-\infty}^{+\infty} g(x) \exp\{-x^2\} dx \approx \sum_i w_i g(x_i)$$

The weights, w_i, and the evaluation points, x_i, can be found in books on numerical integration, e.g., Abramowitz and Stegun (1964).

In general, however, the evaluation of the likelihood can be quite difficult. For the general case,

$$\iint \ldots \int_{\text{dim of u}} \exp(\Sigma_i Y_i(x_i'\beta + z_i'u))\Pi_i(1 + \exp(x_i'\beta + z_i'u))^{-1} dF(u).$$

The dimension of u can get large quickly. For example, in the leaf blight data, the dimension of the integral is larger than 250!

What to do?

Other approaches to ML

Simulation approximations

     Monte Carlo EM

     Monte Carlo Newton-Raphson

     Stochastic approximation

     Importance sampling

Inference using ML would proceed using the usual asymptotic approximations:

ML estimates are asymptotically normal, with SEs coming from second derivatives of the log likelihood.

Tests would be based on the likelihood ratio test, comparing -2loglikelihood for nested models.

Best predicted values would be estimated by calculating E[random effect|data] and plugging in ML or REML estimates. In general, the conditional expected values can't be evaluated in closed form either.

Tests on variances of random effects The usual asymptotic theory breaks down when testing whether the variance components are equal to zero just as with LMMs. For example, in testing whether a single variance component is zero, the large-sample distribution under H_o is a 50:50 mixture of a χ^2_1 and 0.

Summary: ML

+ Known large sample properties

+ Likelihood ratio tests

- Hard to compute for many GLMMs

- Small sample performance needs to be

  assessed for any particular model.

Conditional Inference

A very different approach to random effects is to treat them as nuisance parameters and condition them away.

Classic situation: Matched pairs binary logistic regression.

Example: Do cancer patients get more effective treatment in a major cancer center or a community hospital? Can't directly compare rates. Patients are matched on treatment date, treatment, protocol and other factors. The response is whether or not there is a large shrinkage in their tumor within 90 days.

Data: (1=shrinkage, 0=no shrinkage).

Pair	Cancer Center	Community Hospital
1	1	1
2	1	0
3	1	1
.		
.		
.		
936	0	1

A model:

$Y_{ij} = 1$ for shrinkage and 0 otherwise. i indexes pairs (i=1,2,…N) and j indexes treatment (with j being 1 for a comm hosp. and 2 for a cancer center).

$$Y_{ij} \sim \text{Bernoulli}(p_{ij})$$

$$\text{logit}(p_{ij}) = \mu_i + \beta x_{ij},$$

where $x_{ij} = 0$ for j=1 and 1 for j=2 (cancer center or "treatment" indicator).

μ_i treated as fixed parameters

Maximum likelihood gives

$$\hat{\beta} = 2\log\frac{N_{01}}{N_{10}},$$

where N_{10} is the number of pairs with $Y_{i1}=1$ and $Y_{i2}=0$ and N_{01} is the number of pairs with $Y_{i1}=0$ and $Y_{i2}=1$.

This is perhaps easiest to visualize in a 2×2 format:

	Treatment	
Control	Failure	Success
Failure	N_{00}	N_{01}
Success	N_{10}	N_{11}

The ML estimator is twice the sensible answer.

Remedy? A commonly used approach is that of conditional likelihood.

Basic idea: Derive the sufficient statistics for the μ_i and work with the conditional distribution given those sufficient statistics.

From the form of the density it is clear that the sufficient statistic is $(S_1, S_2, \ldots, S_N, T) = (Y_{1\cdot}, Y_{2\cdot}, \ldots, Y_{N\cdot}, Y_{\cdot 2})$. Since the distribution is discrete, to find the distribution of $\mathbf{S}$ we merely sum over the appropriate values of $\mathbf{Y}$:

$$f_{\mathbf{S},T}(\mathbf{s},t) = \sum_{\mathbf{y}:s_i = y_{i\cdot}, t = y_{\cdot 2}} f_{\mathbf{Y}}(\mathbf{y})$$

$$= C(\mathbf{s},t) \frac{e^{\sum_i \mu_i s_i + \beta t}}{d},$$

where $C(\mathbf{s},t)$ represents the number of combinations of values of $\mathbf{y}$ that satisfy the constraints.

From this it is straightforward to get the marginal distribution of **S**:

$$f_{\mathbf{S}}(\mathbf{s}) = \sum_z f_{\mathbf{S},T}(\mathbf{s},z)$$

$$= \sum_z C(\mathbf{s},z) \frac{e^{\sum_i \mu_i s_i + \beta z}}{d}$$

and the conditional distribution of T given **S**:

$$f_{T|\mathbf{S}}(t|\mathbf{s}) = f_{\mathbf{S},T}(\mathbf{s},t) \Big/ f_{\mathbf{S}}$$

$$= \frac{C(\mathbf{s},t)e^{\beta t}}{\sum_z C(\mathbf{s},z)e^{\beta z}}$$

None of the μ_i remain, as expected. This conditional likelihood can thus be used to estimate β or to form tests or confidence intervals.

For the matched pairs situation the combinatorial coefficient is straightforward to evaluate.

Conditional on $S_i = 0$ we know $Y_{i1}=0$ and $Y_{i2}=0$.
Conditional on $S_i = 2$ we know $Y_{i1}=1$ and $Y_{i2}=1$.

The only remaining randomness involves those pairs for which $S_i = 1$.

Using $r = t - N_{00} - N_{11} =$ number of successes in the discordant pairs, is equivalent to using t.

Then it isn't hard to show that

$C(\mathbf{s},t) =$ number of ways the successes in the N_{10} and N_{10} pairs can be distributed

$$= \binom{N_{10}+N_{01}}{r} = \binom{N'}{r}$$

Illustration: The conditional approach discards the 132+501 = 633 responses which are concordant and bases the analysis on the 303 remaining.

p-value = 2×Pr{X≤146}

where X ~ Binomial(303,1/2).

So p-value = 2(0.283) = 0.566.

Drawbacks to the conditional approach

Recover information from concordant pairs?

Inferences about random effects?

Between versus within "subjects."?

Generalized Estimating Equations

(GEEs)

GEEs are a computationally less demanding method than ML estimation. They are applicable (mainly) to longitudinal data.

Longitudinal data = data collected on a subject on two or more occasions.

Number of occasions is small compared to the number of subjects.

Longitudinal Data

Begin by considering longitudinal data with linear models under normality.

(1) Separate effects that are constant across subjects (β) from those which vary across subjects ($\mathbf{u}_i$).

(2) For the i*th* individual write a linear model conditional on the value of $\mathbf{u}_i$:

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{u}_i + \varepsilon_i$$

$$\varepsilon_i \sim N(\mathbf{0},\mathbf{R}_i)$$

(3) Incorporate subject-to-subject variability by assigning a distribution to $\mathbf{u}_i$:

$$\mathbf{u}_i \sim N(\mathbf{0},\mathbf{D}).$$

Result:　　$\mathbf{Y}_i \sim \text{indep } N(\mathbf{X}_i\beta, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i)$

Longitudinal Data

Example: (Diggle, Liang and Zeger, 1994). Milk was collected from 79 cows on one of three diets: barley, lupins, and a mixture of both. Protein content of the milk was recorded weekly for 19 weeks after the earliest calving.

Constant effects: diet, time

Effects that vary across animals: intercepts

Model for the i*th* cow on diet j, at time t

$$Y_{ijt} = \mu + c_{i(j)} + \alpha_j + f(t) + e_{ijt}$$

$$\mathbf{e}_{ij} \sim N(\mathbf{0}, \mathbf{R}_{i(j)})$$

$$\mathbf{R}_{i(j)}: \ \text{cov}(e_{ijt}, e_{ijt'}) = \sigma_e^2 \exp(-\phi|t-t'|)$$

$$c_{i(j)} \sim N(0, \sigma_c^2)$$

(More generally for awhile):

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \mathbf{e}$$

$$\mathbf{u} \sim N(\mathbf{O}, \mathbf{D})$$

$$\mathbf{e} \sim N(\mathbf{O}, \mathbf{R})$$

So $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V} = \mathbf{ZDZ}' + \mathbf{R})$.

What about using $\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$?

$\hat{\boldsymbol{\beta}}_{OLS}$ is unbiased.

$$E[\hat{\boldsymbol{\beta}}_{OLS}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{Y}]$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

$\hat{\boldsymbol{\beta}}_{OLS}$ is usually fairly efficient.

$$\text{Var}(\hat{\boldsymbol{\beta}}_{OLS}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

(As compared to $\text{Var}(\hat{\boldsymbol{\beta}}_{GLS}) = (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}$)

In fact, with balanced designs, $\hat{\boldsymbol{\beta}}_{OLS} = \hat{\boldsymbol{\beta}}_{GLS}$.

So why not just use $\hat{\boldsymbol{\beta}}_{OLS}$ and standard software?

$$\text{Var}(\hat{\beta}_{OLS}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

but, using standard software,

$\hat{\text{Var}}(\hat{\beta}_{OLS}) = (\mathbf{X}'\mathbf{X})^{-1}\hat{\sigma}^2$, which will often be very wrong. That is, the OLS estimate isn't so bad, but the usual variance estimate is way off.

Going back to the longitudinal data setting the basic idea is, with $Y_i \sim$ independently, to use the "replication" across subjects to get an empirical estimate of the variance. For the longitudinal data setting,

$$\hat{\beta}_{OLS} = (\Sigma_i \mathbf{X}_i' \mathbf{X}_i)^{-1} (\Sigma_i \mathbf{X}_i' \mathbf{Y}_i)$$

$$\text{Var}(\hat{\beta}_{OLS}) = (\Sigma_i \mathbf{X}_i' \mathbf{X}_i)^{-1} (\Sigma_i \mathbf{X}_i' \mathbf{V}_i \mathbf{X}_i) (\Sigma_i \mathbf{X}_i' \mathbf{X}_i)^{-1}$$

which can be estimated by

$$(\Sigma_i \mathbf{X}_i' \mathbf{X}_i)^{-1} (\Sigma_i \mathbf{X}_i' (\mathbf{Y}_i - \hat{\mu}_i)(\mathbf{Y}_i - \hat{\mu}_i)' \mathbf{X}_i)(\Sigma_i \mathbf{X}_i' \mathbf{X}_i)^{-1}$$

For the milk protein data from Diggle, Liang and Zeger (1994), if all the animals had all 19 weeks of data we could just get empirical estimates from the multivariate observations.

With some missing data the previous formula can still be used.

Non-normal data?

GEEs work most easily for models specified on the unconditional distribution. In contrast, we have been specifying models which are conditional on the random effects, u.

For example, for binary data, we could specify:

$$E[Y_{ij}] = p_{ij}$$
$$logit(p_{ij}) = \mathbf{X}_i\boldsymbol{\beta}.$$

Obtain $\hat{\boldsymbol{\beta}}$ by solving the GEE:

$$\sum_{i=1}^{n}\left(\frac{\partial \mathbf{p}_i}{\partial \beta}\right)' W.Var(\mathbf{Y}_i)^{-1}(\mathbf{Y}_i - \mathbf{p}_i) = 0,$$

where *W.Var* indicates a "working" or assumed covariance structure, possibly dependent on unknown parameters.

This has properties similar to the estimating equations for the LMM:

$$\sum_{i=1}^{n}\mathbf{X}_i' \mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}) = 0$$

A big advantage of the GEE approach is the ability to use a "robust" variance estimate.

In such a case the inferences about the mean structure are asymptotically valid, even when the working variance is incorrect.

This offers a useful tool for inference or, at least, model checking.

GEEs are most naturally adapted to marginal models, not the conditional random effects models of GLMMs. But see Zeger, Liang and Albert (1988) see for some results in this direction.

In addition to the drawbacks above relating to marginal models, the GEE approach in particular also has the following drawbacks compared to GLMMs:

- GEEs by themselves do not help to separate out different sources of variation.
- GEEs are not directly a technology for best prediction of random effects. But see Waclawiw and Liang (1993) and Heagerty (1999).
- GEEs are not the best technique for other-than-longitudinal (but correlated) data, either crossed or nested random factors.
- GEEs may be inefficient when the goal is estimation of the variance covariance structure.

Summary: GEEs

Mainly for longitudinal data.

Easiest for marginal models, not random effects models: GEEs by themselves do not help to separate out sources of variation that may be present and do not provide predicted values.

Robust standard errors:
    + Robust
    + Often relatively efficient
    - Estimates many parameters
    - Does not work well when the number of
        time points is large compared to the
        number of subjects
    - Does not work well with missing data

Penalized Quasi-likelihood (PQL)

$\mathbf{Y} \sim$ exponential family with mean μ

$g(\mu) = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} \qquad \mathbf{u} \sim N(\mathbf{0},\mathbf{D})$

$g(\mathbf{y}) \approx g(\mu) + (\mathbf{y}\text{-}\mu)g'(\mu) \equiv \mathbf{z}$

$\qquad = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + (\mathbf{y}\text{-}\mu)g'(\mu)$

$\qquad = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \varepsilon g'(\mu)$

Idea: treat z as a LMM with

$Var(\mathbf{z}) = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R}(g'(\mu))^2$

Use the Mixed Model Equations iteratively to find both

$\hat{\beta}$ and the BLUP of $\mathbf{u}$

Schall (1991) also suggests ways to get approximate SEs.

Summary: PQL

\+ Computationally fairly easy

\+ Works well when the data are

      approximately normal to start with.

\- Does not work well for highly non-normal

      data (e.g., binary).

\- Only for **u** ~ Normal.

Why PQL? See Breslow and Clayton (1993).

Other Approaches

1. Models for specific situations.
   Beta-binomial (Crowder, 1978)
   Poisson-gamma (Abu-Libdeh, et al, 1990)
   Other (Conaway, 1990)

3. Other marginal models
   Liang, Zeger and Qaqish (1992)

4. BLUP estimators
   Engel and Keen (1994)
   McGilchrist (1994,1995)

5. Maximum hierarchical likelihood.
   Lee and Nelder (1996)

More on the beta-binomial

Scenario: A potentially toxic chemical is administered to pregnant rats in the treatment group(s). There is also a control group. The response we record is the presence or absence of a birth defect in animal *k* from litter *j* in group *i*.

$$Y_{ijk} \mid p_{ij} \sim \text{indep. Bernoulli}(p_{ij})$$

$$p_{ij} \sim \text{indep. Beta}(\alpha_i, \beta_i)$$

Hence $Y_{ijk} \sim \text{Bernoulli}(\mu_i)$, where μ_i is given by $E[\,p_{ij}\,] = \alpha_i/(\alpha_i + \beta_i)$.

The joint density of $\mathbf{Y}$ is given by

$$f_{\mathbf{Y}} = \Pi_{i,j} \; f_{\mathbf{Y}_{ij}},$$

where $\mathbf{Y}_{ij} = (Y_{ij1}, Y_{ij2}, \ldots, Y_{ijn_{ij}})'$. Dropping the i and j subscripts,

$$f_{\mathbf{Y}} = \int_0^1 \Pi_k \; p^{Y_k} (1-p)^{(1-Y_k)} \, p^{\alpha-1} (1-p)^{\beta-1} \Big/ B(\alpha,\beta) dp$$

$$= \int_0^1 p^{\alpha+Y.-1} (1-p)^{n-Y.+\beta-1} \Big/ B(\alpha,\beta) dp$$

$$= \frac{B(\alpha+Y., \beta+n-Y.)}{B(\alpha,\beta)}$$

Therefore, the likelihood is given by

$$L = \prod_{ij} \frac{B(\alpha_i + Y_{ij.}, \beta_i + n_{ij} - Y_{ij.})}{B(\alpha_i, \beta_i)}.$$

Extensions? E.g., different doses of the toxic chemical?

Software

<u>Maximum likelihood</u>

Linear Normal Mixed Models: SAS PROC MIXED or SPSS.

Linear Normal Nested Models: MlwiN (http://multilevel.ioe.ac.uk) and HLM (http://www.ssicentral.com/hlm/hlm.htm) fit hierarchical models, using maximum likelihood for normal data and penalized quasi-likelihood for binary and binomial data (see below).

Logit/Probit normal, Ordinal logit: MIXOR program runs on PCs available free from Don Hedeker via the WWW at http://www.uic.edu/~hedeker/mix.html.

Nonlinear normal mixed models: S-Plus functions free from Pinheiro, Bates and Lindstrom at: http://www.stat.wisc.edu/p/stat/ftp/src/NLME/

SAS NLMIXED (new in Version 7.0) can handle random effects for the longitudinal data situation (i.e., data are in clusters).

<u>GEE software</u>

SAS GENMOD allows GEE estimation through its REPEATED statement. SUDAAN and STATA allow GEE estimation for a variety of statistical methods including multinomial logistic regression.

PQL software
GLIMMIX macros available from SAS at
http://ftp.sas.com/techsup/download/stat/glmm800.html

For nested models MlwiN and HLN use PQL and
improvements of PQL.

Bayes software:

BUGS fits a wide variety of Bayesian models and allows
the incorporation of distributions for the parameters. A
description of BUGS (and it can be downloaded from)
http://www.mrc-bsu.cam.ac.uk/bugs/

Case Studies

Case study 1: <u>Breeding Bird Survey</u>. (James, et al, 1996). Counts of number of birds "sighted" has been made each June at thousands of locations across the U.S. and Canada. Many of the locations have been surveyed since the mid 1960s. Responses are summarized by estimating whether the trend in population size is positive within a stratum.

response: increase (yes/no) for species i in stratum j.

distribution: Bernoulli *link:* probit

predictors: species (fixed), stratum (random).

<u>Question:</u> Is destruction of overwintering habitat causing the decline of neo-tropical migrant bird populations on a continent-wide basis?

Breeding Bird Survey (cont)

Model: Y_{ij} = (1/0) increase for species i in stratum j? (Probit-normal)

$$Y_{ij} \sim \text{Bernoulli}(p_{ij})\ i=1,...,26;\ j=1,...,37$$

$$p_{ij} = \Phi(\mu_i + s_j)\,, \qquad s_j \sim N(0, \sigma_s^2)$$

Data layout

		Stratum						
Species	1	...	10	11	12	13	...	37
1			0			0		
2								
3								
4	0							
.					0	0		
26			1	1	1	0		0

blank = species not present (about 2/3)

1 = increase

0 = decrease

ML Estimates:

$$\hat{\mu}_1 = -0.66, \hat{\mu}_2 = 0.26, ..., \hat{\mu}_{26} = -1.28$$

$$\hat{\sigma}_s^2 = 0.45$$

Interpretations:

$$\Phi(0.26) = 0.60$$

$$\hat{\sigma}_s^2/(\hat{\sigma}_s^2 + 1) = 0.31$$

Test of $\sigma_s^2 = 0$:

diff in -2loglik = 11.88

compare to a $\frac{1}{2}\chi_1^2$

Estimated best predicted values: $E[s_j|Y]$

e.g., $E[s_{23}|Y] = -1.10$

$$\Phi(-1.10) = 0.14$$

Case Study 2: <u>Progabide and Seizures</u> (Diggle, Liang and Zeger, 1994). Epileptics were randomly allocated to a placebo group or an drug (Progabide) group. The number of seizures was recorded for a baseline period of 8 weeks and during consecutive two-week periods for 4 periods after beginning treatment. Is the drug effective at reducing the number of seizures?

| | Number of seizures | | | | | |
Patient	Base-line	Period 1	Period 2	Period 3	Period 4	Trt
1	11	5	3	3	3	0
2	11	3	5	3	3	0
3	6	2	4	0	5	0
4	8	4	4	1	4	0
.	.	.	.	.	.	.
57	13	0	0	0	0	1
58	12	1	4	3	2	1

response: number of seizures for individual i at

time j=1,2,3,4,5

distribution: Poisson

predictors: period, treatment (both fixed),
individual, individual × treatment (?) (both
random).

The baseline period is 8 weeks long, whereas the
observation periods are only 2 weeks long.

Question: Does Progabide reduce the frequency
of seizures?

Model:

Y_{ij} = count for subject i at time j

t_{ij} = time (in weeks) for the observation period for subject i at time j (either 8 or 2 weeks).

$$Y_{ij}|\lambda_{ij} \sim \text{indep. Poisson}(\lambda_{ij})$$

$$\ln(\lambda_{ij}) = \mu + s_i + \beta_1 TIME_{ij} + \beta_2 TRT_{ij} +$$

$$\beta_3 TIME_{ij} \times TRT_i + \ln(t_{ij})$$

$$s_i \sim N(0, \sigma_s^2)$$

$TIME_{ij} = 1$ if the observation is post baseline and 0 otherwise.

Mainly interested in β_3.

How to estimate this model?

SAS Programs for the Progabide data

```
data thall;
input id y visit trt bline age;
cards;
104 5 1 0 11 31
104 3 2 0 11 31
104 3 3 0 11 31
103 0 4 1 19 20
...
232 0 4 1 13 36
236 1 1 1 12 37
236 4 2 1 12 37
236 3 3 1 12 37
236 2 4 1 12 37
;

data new;
   set thall (drop=age);
   output;
   if visit=1 then do; y=bline; visit=0; output; end;
   run;

proc sort;
   by id visit;
   run;

data new3;
   set new;
   if id ne 207;
   if visit=0 then do; time=0; ltime=log(8); end;
   else do; time=1; ltime=log(2); end;
   run;

proc nlmixed data=new3 qpoints=20;
   parms  mu=1 b1=0 b2=0 b3=0 sig1=0.1;
   eta=mu+b1*time+b2*trt+b3*time*trt+u1+ltime;
   lam=exp(eta);
   model y~Poisson(lam);
   random u1~Normal(0,sig1) subject=id;
   run;

proc nlmixed data=new3 qpoints=20;
   parms  mu=1 b1=0 b2=0 b3=0 sig1=0.1 cov=0.05 sig2=0.1;
   eta=mu+b1*time+b2*trt+b3*time*trt+u1+u2*time+ltime;
   lam=exp(eta);
   model y~Poisson(lam);
   random u1 u2~Normal([0, 0],[sig1, cov, sig2]) subject=id;
   run;

proc genmod data=new3;
   class id;
   model y= time trt time*trt / d=poisson offset=ltime;
   repeated subject=id / corrw covb type=exch;
   run;
```

SAS output

```
                      The NLMIXED Procedure

                        Specifications

         Data Set                             WORK.NEW3
         Dependent Variable                   y
         Distribution for Dependent Variable  Poisson
         Random Effects                       u1
         Distribution for Random Effects      Normal
         Subject Variable                     id
         Optimization Technique               Dual Quasi-Newton
         Estimation Method                    Adaptive Gaussian
                                              Quadrature


                          Dimensions

         Observations Used              290
         Observations Not Used            0
         Total Observations             290
         Subjects                        58
```

```
                Max Obs Per Subject                    5
                Parameters                             5
                Quadrature Points                     20


                            Parameters

        mu          b1          b2          b3        sig1     NegLogLike

         1           0           0           0         0.1    1015.04066


                            Iterations

     Iter     Calls     NegLogLike        Diff      MaxGrad       Slope

        1         3       955.8817    59.15896     57.97191    -7021.37
        2         5     952.178211     3.70349     55.02552     -34.165
        3         6     948.800206    3.378005     13.73927    -29.1799
        4         7     948.525761    0.274445     2.674937    -0.54631
        5         9     948.511099    0.014661     1.814969    -0.00427
        6        10     948.486503    0.024596     0.610316    -0.02313
        7        12     948.483431    0.003072     0.070732    -0.00642
        8        14     948.483277    0.000154     0.052537    -0.00008
        9        16     948.483246    0.000031     0.002884    -0.00005
       10        18     948.483246    3.612E-8     0.000061    -7.42E-8
```

The NLMIXED Procedure

NOTE: GCONV convergence criterion satisfied.

Fitting Information

-2 Log Likelihood	1897.0
AIC (smaller is better)	1907.0
BIC (smaller is better)	1917.3
Log Likelihood	-948.5
AIC (larger is better)	-953.5
BIC (larger is better)	-958.6

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
mu	1.0359	0.1415	57	7.32	<.0001	0.05	0.7526	1.3192	0.000061
b1	0.1108	0.04689	57	2.36	0.0216	0.05	0.01691	0.2047	-0.00002
b2	-0.01049	0.1968	57	-0.05	0.9577	0.05	-0.4047	0.3837	0.000044
b3	-0.3016	0.06975	57	-4.32	<.0001	0.05	-0.4413	-0.1619	0.000041
sig1	0.5167	0.1013	57	5.10	<.0001	0.05	0.3139	0.7196	-0.00001

The NLMIXED Procedure

Specifications

Data Set	WORK.NEW3
Dependent Variable	y
Distribution for Dependent Variable	Poisson
Random Effects	u1 u2
Distribution for Random Effects	Normal
Subject Variable	id
Optimization Technique	Dual Quasi-Newton
Estimation Method	Adaptive Gaussian Quadrature

Dimensions

Observations Used	290
Observations Not Used	0
Total Observations	290
Subjects	58
Max Obs Per Subject	5
Parameters	7
Quadrature Points	20

Parameters

mu	b1	b2	b3	sig1	cov	sig2	NegLogLike
1	0	0	0	0.1	0.05	0.1	952.625769

Iterations

Iter	Calls	NegLogLike	Diff	MaxGrad	Slope
1	3	912.039756	40.58601	264.2311	-7563.74
2	4	898.775177	13.26458	18.03295	-979.652
3	5	896.722486	2.052691	13.29285	-4.27047
4	7	895.336636	1.38585	12.48232	-1.57163
5	8	894.589275	0.747361	10.88548	-2.00732
6	9	894.365239	0.224036	12.92942	-0.83035
7	11	893.837266	0.527973	8.323319	-1.0606
8	13	893.468515	0.36875	10.17836	-0.16147
9	15	893.346311	0.122204	10.68736	-0.15449
10	16	893.139226	0.207085	5.499622	-0.11613
11	18	893.053203	0.086023	0.895064	-0.18358
12	20	893.046781	0.006421	0.235226	-0.01462
13	22	893.045961	0.000821	0.13539	-0.00128

```
                        The NLMIXED Procedure

                             Iterations

        Iter    Calls    NegLogLike       Diff    MaxGrad      Slope

          14       24    893.045484    0.000477   0.121494   -0.00047
          15       26    893.045332    0.000152   0.029424   -0.00026
          16       28     893.04533    1.603E-6    0.00084   -3.17E-6


            NOTE: GCONV convergence criterion satisfied.


                        Fitting Information

            -2 Log Likelihood                    1786.1
            AIC (smaller is better)              1800.1
            BIC (smaller is better)              1814.5
            Log Likelihood                       -893.0
            AIC (larger is better)               -900.0
            BIC (larger is better)               -907.3

                        Parameter Estimates

                         Standard          t
   Parameter   Estimate    Error   DF   Value  Pr > |t|  Alpha    Lower     Upper   Gradient

      mu         1.0696    0.1343   56   7.96   <.0001    0.05   0.8005    1.3387   -0.0006
      b1       0.005870    0.1070   56   0.05   0.9564    0.05  -0.2085    0.2202    0.000209
      b2       -0.00970    0.1860   56  -0.05   0.9586    0.05  -0.3823    0.3629   -0.0005
      b3        -0.3471    0.1489   56  -2.33   0.0233    0.05  -0.6453  -0.04888   -0.00024
      sig1       0.4528   0.09354   56   4.84   <.0001    0.05   0.2654    0.6402    0.000059
      cov        0.01725  0.05287   56   0.33   0.7455    0.05  -0.08867   0.1232   -0.00084
      sig2       0.2161   0.05864   56   3.69   0.0005    0.05   0.09862   0.3336   -0.00047
```

```
                      The GENMOD Procedure

                      Model Information

             Data Set              WORK.NEW3
             Distribution           Poisson
             Link Function              Log
             Dependent Variable           y
             Offset Variable          ltime
             Observations Used          290


                  Class Level Information

   Class     Levels   Values

   id            58   101 102 103 104 106 107 108 110 111 112 113 114
                      116 117 118 121 122 123 124 126 128 129 130 135
                      137 139 141 143 145 147 201 202 203 204 205 206
                      208 209 210 211 213 214 215 217 218 219 220 221
                      222 225 226 227 228 230 232 234 236 238


                      Parameter Information

                  Parameter        Effect

                  Prm1             Intercept
                  Prm2             time
                  Prm3             trt
                  Prm4             time*trt


               Criteria For Assessing Goodness Of Fit

        Criterion              DF       Value      Value/DF

        Deviance              286    2413.0245       8.4371
        Scaled Deviance       286    2413.0245       8.4371
        Pearson Chi-Square    286    3015.1555      10.5425
        Scaled Pearson X2     286    3015.1555      10.5425
        Log Likelihood               5631.7547


   Algorithm converged.
```

The GENMOD Procedure

Analysis Of Initial Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
				Lower	Upper		
Intercept	1	1.3476	0.0341	1.2809	1.4144	1565.44	<.0001
time	1	0.1108	0.0469	0.0189	0.2027	5.58	0.0181
trt	1	-0.1080	0.0486	-0.2034	-0.0127	4.93	0.0264
time*trt	1	-0.3016	0.0697	-0.4383	-0.1649	18.70	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

GEE Model Information

Correlation Structure	Exchangeable
Subject Effect	id (58 levels)
Number of Clusters	58
Correlation Matrix Dimension	5
Maximum Cluster Size	5
Minimum Cluster Size	5

Covariance Matrix (Model-Based)

	Prm1	Prm2	Prm3	Prm4
Prm1	0.01223	0.001520	-0.01223	-0.001520
Prm2	0.001520	0.01519	-0.001520	-0.01519
Prm3	-0.01223	-0.001520	0.02495	0.005427
Prm4	-0.001520	-0.01519	0.005427	0.03748

Covariance Matrix (Empirical)

	Prm1	Prm2	Prm3	Prm4
Prm1	0.02476	-0.001152	-0.02476	0.001152
Prm2	-0.001152	0.01348	0.001152	-0.01348
Prm3	-0.02476	0.001152	0.03751	-0.002999
Prm4	0.001152	-0.01348	-0.002999	0.02931

Algorithm converged.

```
                       The GENMOD Procedure

                  Working Correlation Matrix

              Col1        Col2        Col3        Col4        Col5

Row1        1.0000      0.5941      0.5941      0.5941      0.5941
Row2        0.5941      1.0000      0.5941      0.5941      0.5941
Row3        0.5941      0.5941      1.0000      0.5941      0.5941
Row4        0.5941      0.5941      0.5941      1.0000      0.5941
Row5        0.5941      0.5941      0.5941      0.5941      1.0000


                 Analysis Of GEE Parameter Estimates
                  Empirical Standard Error Estimates

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

          Intercept   1.3476  0.1574   1.0392   1.6560     8.56  <.0001
          time        0.1108  0.1161  -0.1168   0.3383     0.95  0.3399
          trt        -0.1080  0.1937  -0.4876   0.2716    -0.56  0.5770
          time*trt   -0.3016  0.1712  -0.6371   0.0339    -1.76  0.0781
```

Parameter estimates and SEs (in parenthesis) for the Progabide data.

	Estimation Method		
Variable	MLE[1]	PQL[2]	GEE[3]
Intercept	1.03	1.00	1.35
	(0.14)	(0.14)	(0.16)
TRT	-0.01	-0.009	-0.11
	(0.20)	(0.19)	(0.19)
TIME	0.11	0.11	0.11
	(0.05)	(0.05)	(0.12)
TIME×TRT	-0.30	-0.30	-0.30
	(0.07)	(0.07)	(0.17)
σ_s^2	$\hat{\sigma}_s^2=0.52$	$\hat{\sigma}_s^2=0.53$	$\hat{\rho}=0.60$
	(0.10)	(0.10)	

[1]SAS Proc NLMIXED

[2]From Diggle, Liang, and Zeger (1994, p.188)

[3]SAS Proc GENMOD

Case Study 3: <u>Potomac River Fever in Horses:</u> (Atwill, et al, 1996) Potomac River Fever (equine monocytic ehrlichiosis) is a blood-borne rickettsial disease whose transmission mechanism is unknown. Both arthropod (e.g. blackfly) and direct oral transmission have been suspected but not verified. Identification of risk factors of horses in New York State might give clues to the spread of this disease and help with reducing its frequency.

511 farms were studied, each with several social groups of horses, for a total of 2,587 horses.

response: seropositive (yes/no) response for horse k in social group j at farm i.

distribution: Bernoulli *link:* logit

predictors: Frequency stall cleaned, Frequency fly spray applied, Breed, Sex, ...(fixed), Farm and Social group (farm) (random).

<u>Questions:</u>   Transmission mechanism of Potomac River Fever?

<u>Model</u>: $Y_{ijk} = $ infection for horse k in social
    group j on farm i.

    $Y_{ijk} \sim$ Bernoulli(p_{ijk})

    logit(p_{ijk}) = $\mu + s_{j(i)} + f_i + $ fixed effects,

        $s_{j(i)} \sim N(0, \sigma^2_{group(farm)})$

        $f_i \sim N(0, \sigma^2_{farm})$

<u>Analysis</u>: Focus on the random factors. The estimated variances of the random effects were:

$$\hat{\sigma}^2_{farm} = 1.26$$
$$\hat{\sigma}^2_{group(farm)} = 0$$

So the difference in loglikelihood for testing $\sigma^2_{group(farm)} = 0$ is zero and hence not statistically significant when compared to a $\frac{1}{2}\chi^2_1$.

<u>Implications</u>: There is a strong correlation among horses within a farm on the logit scale ($0.32 = \hat{\sigma}^2_{farm}/(\hat{\sigma}^2_{farm}+\sigma^2_{logistic})$)), but no correlation within social groups. This suggests the disease is not transmitted directly from horse to horse, but instead is related to environmental or management factors operating at a farm scale.

Case study 4: <u>Chestnut Leaf Blight</u>. Recall the situation: Viruses spread between fungal individuals when they come in contact and fuse together. A major obstacle in spreading this virus and thus controlling the disease is that different isolates of the fungus cannot necessarily transfer the virus to one another.

To estimate the effects of these genes, they have paired numerous isolates which differ on the first gene only, the second gene only, the first and the second gene, etc. For each combination of isolates they have averaged about 30 attempts and record a binary response of whether or not the attempt succeeded in transmitting the virus.

Questions of interest include whether pre-identified genes actually do have an influence on transmission of the virus (and if so, to what degree), whether there are other, as yet unidentified, genes which might affect transmission, and whether transmission is symmetric. By symmetry of transmission we mean the following: suppose the infected fungus is type b at the locus for the first gene and the non-infected isolate (which we are trying to infect) is type B. The two isolates are the same at the other five loci. Is the probability of transmission the same as when using a type B to try to infect a type b?

Model:

$Y_i = 1$ if virus is transmitted, 0 otherwise
$Y_i \sim$ indep. Bernoulli(p_i)

$$p_i = \Phi(\mu + \Sigma_j \, \beta_j MCH_{ij} + \Sigma_j \gamma_j ASY_{ij}),$$

where $MCH_{ij} = 1$ if there is a mismatch at locus j for pairing i and 0 otherwise, and $ASY_{ij} = \frac{1}{2}$ if there is a mismatch at locus j in pairing i with a b donor, $-\frac{1}{2}$ if there is a mismatch at locus j pairing i with a B donor and 0 if there is no mismatch.

β_j = effect of a mismatch on gene j.

γ_j = asymmetry effect.

   = difference between a mismatch with a donor type b and type B.

Question: Is there asymmetric transmission?

maximized log likelihood of the model:
  log l = -955.303
  with 13 parameters

maximized log likelihood of the model with all the γ_i set equal to zero:
  log l = -1116.639
  with 7 parameters

Likelihood ratio test:

Difference is 1116.639 - 955.303 = 161.336

p-value = $P\{\chi^2_6 \geq 2*161.336\} \approx 0$

Threshold model

A common model in genetics for describing the presence or absence of a trait is the threshold model. This arises from assuming that a large number of genes each have a small and additive effect and when the cumulative effect exceeds a threshold of zero the trait is present in an individual.

$Y = 1$ if trait is present, and 0 otherwise.
$\mathbf{x}'\boldsymbol{\beta}$ = either genetic or non-genetic fixed effects.
ε = the genetic effect not captured in $\mathbf{x}'\boldsymbol{\beta}$.

Appealing to the central limit theorem gives the probit model:

$$P\{Y=1\} = P\{\mathbf{x}'\boldsymbol{\beta} + \varepsilon > 0\}$$
$$= P\{-\varepsilon < \mathbf{x}'\boldsymbol{\beta}\} = \Phi(\mathbf{x}'\boldsymbol{\beta})$$

Different isolates of the fungus are used which may differ with regard to genes other than the six pre-identified.

We might model their effects as being selected from a normal distribution.

Y_{ijk} = kth observation from an attempted infection from the ith isolate (the donor) to the jth isolate (the recipient).

$\mathbf{x}_{ijk}$ = vector of covariates for Y_{ijk}

A reasonable model might then be:

$$P\{Y_{ijk} = 1 \mid \mathbf{u}\} = P\{\mathbf{x}'_{ijk}\boldsymbol{\beta} + u_{1i} + u_{2j} + \varepsilon_{ijk} > 0\},$$

where u_{1i} represent the (random) effects of the donor isolate and u_{2j} represent the (random) effects of the recipient isolate.

This gives

$$P\{Y_{ijk} = 1 \mid \mathbf{u}\} = \Phi(\mathbf{x}'_{ijk}\boldsymbol{\beta} + u_{1i} + u_{2j})$$

Consequences of introducing random factors

On the mean

$$E\left[Y_{ijk}\right] = E\left[E[Y_{ijk} \mid \mathbf{u}]\right] = \Phi(\mathbf{x}'_{ijk}\boldsymbol{\beta} + u_{1i} + u_{2j}),$$

or, using the threshold representation:

$$E\left[Y_{ijk}\right] = E\left[P\{\mathbf{x}'_{ijk}\boldsymbol{\beta} + u_{1i} + u_{2j} + \varepsilon_{ijk} > 0 \mid \mathbf{u}\}\right]$$

$$= P\{\mathbf{x}'_{ijk}\boldsymbol{\beta} + u_{1i} + u_{2j} + \varepsilon_{ijk} > 0\}$$

$$= P\{-(u_{1i} + u_{2j} + \varepsilon_{ijk}) < \mathbf{x}'_{ijk}\boldsymbol{\beta}\}$$

$$= P\{W < \mathbf{x}'_{ijk}\boldsymbol{\beta}\},$$

where $W \sim \mathrm{N}(\,0,\,1 + \sigma_1^2 + \sigma_2^2\,)$. So

$$E\left[Y_{ijk}\right] = \Phi\left(\mathbf{x}'_{ijk}\boldsymbol{\beta} \Big/ \sqrt{1 + \sigma_1^2 + \sigma_2^2}\right)$$

$$= \Phi\left(\mathbf{x}'_{ijk}\boldsymbol{\beta}^*\right)$$

On the variance-covariance structure

For example, for two observations with the same donor and recipient isolate:

$$E\left[Y_{ijk}Y_{ijl}\right] =$$

$$\int_{-\infty}^{+\infty} \Phi\left(\mathbf{x}'_{ijk}\boldsymbol{\beta} + \sigma z\right)\Phi\left(\mathbf{x}'_{ijl}\boldsymbol{\beta} + \sigma z\right)\exp(-z^2/2)/\sqrt{2\pi}\,dz,$$

where $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$.

A correlation is therefore induced between responses which share one or more random effects.

Likelihood:

The conditional density of **Y** given **u** is
$$f_{\mathbf{Y}|\mathbf{u}} =$$

$$\Pi \Phi(\mathbf{x}'_{ijk}\boldsymbol{\beta} + u_{1i} + u_{2j})^{y_{ijk}} [1 - \Phi(\mathbf{x}'_{ijk}\boldsymbol{\beta} + u_{1i} + u_{2j})]^{1-y_{ijk}},$$

so that the likelihood is given by
$$L = \int \cdots \int f_{\mathbf{Y}|\mathbf{u}} f_{\mathbf{u}} d\mathbf{u},$$

which, for this example, is a 256-dimensional integral.

<u>Question of interest</u>: Are there other genes causing incompatibility?

If there are no other genes affecting the transmission of the virus, then all isolates with a given set of fixed effects will behave the same.

$\Rightarrow$ H$_0$: $\sigma_1^2 = 0, \sigma_2^2 = 0$

Suppose we reject H$_0$. How could we go about finding the genes that control incompatibility? We might look at the isolates that have the most extreme values of u_{1i} or u_{2j}.

Extreme values of u_{1i} or u_{2j}: Want predicted values of the u_{1i} and u_{2j}.

$$\text{best predictor} = \tilde{u}_{1i} = E\left[u_{1i} \mid \mathbf{Y}\right]$$

Two problems

- Depends on unknown parameters
- $E\left[u_{1i} \mid \mathbf{Y}\right] = \int u_{1i} f_{\mathbf{u}\mid\mathbf{Y}} d\mathbf{u}$ and

$$f_{\mathbf{u}\mid\mathbf{Y}} = f_{\mathbf{Y},\mathbf{u}} \big/ f_{\mathbf{Y}}$$

Research

Some selected topics in research.

1. <u>Computing maximum likelihood estimates.</u>
McCulloch (1994) - uses Gibbs sampler
McCulloch (1997) - uses Metropolis
Booth and Hobert (1999) – Indep. sampler
Geyer (1994) - Simulated ML
Geyer and Thompson (1992) - Simulated ML
Econometrics literature (Borsch-Supan and
     Hajivassiliou, 1993)
Casella and Berger (1995) - Another method of
     simulating to find ML estimates
Ruppert, et al (1984) - Stochastic approximation

2. <u>PQL, Laplace approximations</u>
Gilmour, Anderson and Rae (1984)
Schall (1991)
Breslow and Clayton (1994)
Breslow and Lin (1995)
Lin and Breslow (1996)
Wolfinger (1994)

3. <u>Bayes estimates</u>
Gilks, et al (1993)
Zeger and Karim (1991)
(But) Natarajan and McCulloch (1995)

4. <u>GEEs</u>
Zeger and Liang (1986)
Liang and Zeger (1986)
(But) Fitzmaurice (1995), Lipsitz, et al (1994)

5. <u>Other</u>
Engel and Keen (1994)
Kuk (1995)
McGilchrist (1994, 1995)
Heagerty and Lele (1998)
Drum and McCullagh (1993)
 (1999)

SUMMARY

The Good News:

GLMMs can handle
    Non-normal data
    Nonlinear responses
    Random effects covariance structure

Can be used to
    Incorporate correlations in models
    Model the correlation structure
    Identify sensitive subjects
    Handle heterogeneous variances

Modelling process
    1. Distribution of the data?
    2. What is to be modelled?
    3. Factors?
    4. Fixed or random?

Software is available for linear and nonlinear normal models, some GLMs with normal random effects and for GEE estimation.

The not-so-Good News:

Computing methods for much of the class of GLMMs is an area of active research. Advances are being made in ML estimation, PQL, GEEs and Bayes methods.

General purpose software is still developing.

Tests and confidence intervals are asymptotic and approximate.