The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

1863-9

Advanced School and Conference on Statistics and Applied Probability in Life Sciences

24 September - 12 October, 2007

Generalized linear models I

Peter McCullagh

Department of Statistics
University of Chicago
Chicago IL 60637, USA

Generalized linear models I

Peter McCullagh

Department of Statistics
University of Chicago

Trieste, October 2007

Outline

Conventional set-up for regression

Study of dependence of response Y on covariates x

Set of units: $\mathcal{U} = \{u, u', \ldots\}$ subjects, patients, plots,

Covariate $x(u), x(u'), \ldots$ (non-random, vector-valued)
- classification variables: sex, age, soil type,...
- block factors: clinical centre, herd (veterinary applications)
- treatment factor: medication used, variety planted,..

Response: $Y(u)$: yield, survival,... measured on the units
Sample: a finite non-random ordered subset U of the units:

Observation: (data)
- $Y \equiv Y[U] = (Y(u_1), \ldots, Y(u_n))$ on the sampled units
- $X \equiv x[U] = (x(u_1), \ldots, x(u_n))$ on the sampled units

Q1: How does the joint distribution of Y depend on x?
Q2: What effect does treatment have on the response?
Q3: What is the conditional distribution of $Y(u')$ given the data?

Simplifying assumption I: Exchangeability

Exchangeability modulo covariates:

$x[U] = x[U']$ implies $Y[U] \sim Y[U']$ for all $U, U' \subset \mathcal{U}$

where $X = x[U]$ is the covariate value for sample U.

Rationale:

Equivalent sets of units have the same distribution

Implications:

(i) Treatment applied to u' has no effect on $Y(u)$
(ii) Variables other than x are irrelevant
(iii) Relationships among units have no effect on distn

Simplifying assumption II: Independence

Independence: $Y(u_1), \ldots,$ are independent
(Response values for distinct units are independent)

Implications:
- Relationships among units are irrelevant
  - . . . unless encoded in x
- Geographical proximity or genetic relationships

Independence assumption is overused and often taken for granted
Can be relaxed by introducing variance components...

Standard Gaussian model

$U = \{u_1, \ldots, u_n\}$ are the n sampled units in order
$Y = Y[u]$ is the response vector with n components
Covariate $x(u) = (x_1(u), \ldots, x_p(u))$ has p components
Covariate matrix $X = x[U]$ is of order $n \times p$

$$\begin{aligned} Y &\sim N(\mu = X\beta, \sigma^2 I_n) \\ E(Y(u)) &= \beta_1 x_1(u) + \cdots + \beta_p x_p(u) \\ \operatorname{cov}(Y(u), Y(u')) &= \sigma^2 \delta_{u,u'} \end{aligned}$$

Interpretation of β_p:

> β_p is the change in $E(Y)$ per unit change in x_p when $x_1, \ldots, x_{p-1}$ are held fixed

Limitations...

Parameter estimation

Maximum likelihood: Probability density at y in $\mathcal{R}^n$ is

$$(2\pi)^{-n/2}\sigma^{-n}\exp(-(y-\mu)'(y-\mu)/(2\sigma^2))$$

Log likelihood is

$$-n\log\sigma - \sum(y_i-\mu_i)^2/(2\sigma^2)$$

Max likelihood: Choose $\hat{\beta}$ by minimizing $(Y-X\beta)'(Y-X\beta)$
Gives $X'(Y-X\hat{\beta}) = 0$ or $X'X\hat{\beta} = X'Y$
or $\hat{\beta} = (X'X)^{-1}X'Y$
or $\hat{\mu} = X\hat{\beta} = X(X'X)^{-1}X'Y = PY, \quad P^2 = P$

$E(\hat{\beta}) = (X'X)^{-1}X'E(Y) = (X'X)^{-1}X'X\beta = \beta$
$\mathrm{cov}(\hat{\beta}) = (X'X)^{-1}X'\,\mathrm{cov}(Y)X(X'X)^{-1} = \sigma^2(X'X)^{-1}$

Estimate of variance
$s^2 = (Y-\hat{\mu})'(Y-\hat{\mu})/(n-p)$

Example: vitiman C decay

Ascorbic acid concentration of snap beans

	Weeks of storage				
Temp °F	2	4	6	8	Total
0	45	47	46	46	184
10	45	43	41	37	166
20	34	28	21	16	99

(Each value is a sum for three packages)

Same data in spreadsheet format

Conc	time	temp
45	2	0
45	2	10
34	2	20
47	4	0
43	4	10
28	4	20
⋮	⋮	⋮

Vitamin C (continued)

Units $\mathcal{U}$: packages of beans (an infinite set)
Sampled units U: 36 packages with values as given
Data: Y (vitamin C content of sampled packages)

- x_1 storage temperature
- x_2 storage time

Naive linear model

$$E(Y(u)) = \beta_0 + \beta_1 x_1(u) + \beta_2 x_2(u)$$

Fitted coefficients:

Parameter	Estimate	s.e.
β_0	55.125	3.76
time	−1.42	0.61
temp	−1.06	0.17
σ	4.75	1.12

Residual SS = 203.38 on 9 d.f.

Model checks

Criticism: I:

Dependence on temperature is not linear:

Plot the data or the residuals against temp

Remedy: include a different coefficient for each temperature

Modified linear model

$$E(Y(u)) = \beta_0 I_{x_1=0} + \beta_{10} I_{x_1=10} + \beta_{20} I_{x_1=20} + \beta_2 x_2(u)$$

Parameter	Estimate	s.e.
temp0	53.08	2.93
temp10	48.58	2.93
temp20	31.83	2.93
time	−1.42	0.46
σ	4.75	1.12

Residual SS = 103.33 on 8 d.f. (down from 203.38)

Further model checks

Criticism II:

fitted value at time zero depends on the temperature

Remedy: Make the slope (decay rate) depend on temperature

Modified linear model

$$E(Y(u)) = \alpha + \beta_0 tI_{x_1=0} + \beta_{10} tI_{x_1=10} + \beta_{20} tI_{x_1=20}$$

Fitted coefficients:

Parameter	Estimate	s.e.
α	44.50	1.28
β_0	0.27	0.28
β_{10}	−0.72	0.28
β_{20}	−3.80	0.28
σ	3.25	0.81

Residual SS = 26.03 on 8 d.f. (down from 203 and 103)

Further model checks

Criticism III:

the values cannot increase over time

Exponential decay more plausible than linear decay

Remedy: Use a non-linear (Gaussian generalized linear) model

Modified model (generalized linear)

$$\log E(Y(u)) = \alpha + \beta_0 tl_{x_1=0} + \beta_{10} tl_{x_1=10} + \beta_{20} tl_{x_1=20}$$

Parameter	Estimate	s.e.
α	3.84	0.019
β_0	−0.001	0.004
β_{10}	−0.024	0.004
β_{20}	−0.133	0.006
σ	1.12	0.28

Residual SS = 9.00 on 8 d.f. (down from 203, 103 and 26)

Model formulae

Building blocks for subspaces of $\mathcal{R}^n$:

- constant vector 1: all components are equal
- covariate vectors $x_1, x_2, \ldots$ in $\mathcal{R}^n$
- factors $A, B, \ldots$

A factor is a list of levels, e.g. row or column or temperature
A factor is also the subspace spanned by the levels
$1 \subset A$
$\dim(A) =$ number of levels

Binary operators:
$+$ as in $x_1 + x_2$ or $A + x$ or $A + B$ (vector span)
: or . or $*$: Interaction, tensor product
$x + x = x, \quad A : A = A, \quad A : (B + C) = A : B + A : C$

Model formulae for subspaces of $\mathcal{R}^n$

$1 + x_1 + x_2$: $\beta_0 + \beta_1 x_(u) + \beta_2 x_2(u)$
$A + x$: $E(Y(u)) = \alpha_{A(u)} + \beta x(u)$
$A + B$: $E(Y(u)) = \alpha_{A(u)} + \beta_{B(u)}$
$1 + A : x$: $E(Y(u)) = \alpha + \beta_{A(u)} x$
$A : B$: $E(Y(u)) = \beta_{A(u),B(u)}$

Factorial models

One factor A:

```
1,   A
```

Two factors A, B:

```
1,  , A,  , B,  , A+B,  , A*B
```

Three factors A, B, C:

```
1,  , A,B,C,  , A+B[3],  A*B[3],  A+B+C,
A*B+C[3],  A*B+A*C[3],  A*B+A*C+B*C,  A*B*C
```

How many factorial models for k factors?

Free distributive lattice on k generators $A_1, \ldots, A_k$
operators $+$ and $*$ such that $A + A = A$ and $A * A = A$

Factorial models: Latin square

Time taken and saw used

		Team Number					
Spc	Bark	1	2	3	4	5	6
spruce	0	6.4, 6	10.9, 5	9.8, 4	7.5, 2	4.6, 1	4.1, 3
pine	0	6.8, 2	6.2, 3	7.9, 5	6.0, 1	4.0, 4	4.2, 6
larch	0	12.7, 5	13.4, 1	12.5, 2	7.3, 3	6.1, 6	7.4, 4
spruce	1	8.8, 3	10.2, 4	12.5, 1	8.6, 6	6.1, 5	5.6, 2
pine	1	7.4, 4	10.0, 2	8.3, 6	6.4, 5	4.3, 3	5.6, 1
larch	1	13.1, 1	12.0, 6	12.0, 3	11.3, 4	6.1, 2	9.7, 5

Three species, spruce, pine and larch
Bark present or not
Six saws of three types 1&4, 2&5, 3&6

```
nobs <- 36
data <- matrix( c(
+  6.4, 6, 10.9, 5,  9.8, 4,  7.5, 2, 4.6, 1, 4.1, 3,
+  6.8, 2,  6.2, 3,  7.9, 5,  6.0, 1, 4.0, 4, 4.2, 6,
+ 12.7, 5, 13.4, 1, 12.5, 2,  7.3, 3, 6.1, 6, 7.4, 4,
+  8.8, 3, 10.2, 4, 12.5, 1,  8.6, 6, 6.1, 5, 5.6, 2,
+  7.4, 4, 10.0, 2,  8.3, 6,  6.4, 5, 4.3, 3, 5.6, 1,
+ 13.1, 1, 12.0, 6, 12.0, 3, 11.3, 4, 6.1, 2, 9.7, 5), nobs, 2, byrow=T)
y <- data[,1]                    # time taken to complete task
team <- gl(6, 1, nobs)           # six teams, one per column
species <- gl(3, 6, nobs)        # spruce pine and larch
bark <- gl(2, 18, nobs)          # Bark present or not
row <- gl(6, 6, nobs)            # equal to species*bark
saw <- as.factor(data[,2])       # distinct saws (3 duplicate pairs)
stype <- as.factor(data[,2] - 3*(data[,2] > 3)) # saw type

# Gaussian factorial models
y <- log(data[,1])      #transform to log scale
fit0 <- glm(y~row+team+saw, family=gaussian(link=identity))
fit1 <- glm(y~species:bark+team+saw, family=gaussian(link=identity))
fit2 <- glm(y~species+bark+team+saw, family=gaussian(link=identity))
summary(fit0)

Call:
glm(formula = y ~ row + team + saw, family = gaussian(link = identity))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.15957  -0.06465   0.01093   0.05275   0.11339

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  2.125516   0.067189  31.635  < 2e-16 ***
row2        -0.176319   0.058187  -3.030 0.006610 **
row3         0.330140   0.058187   5.674 1.49e-05 ***
row4         0.204383   0.058187   3.513 0.002191 **
row5        -0.003565   0.058187  -0.061 0.951756
row6         0.428042   0.058187   7.356 4.15e-07 ***
team2        0.142463   0.058187   2.448 0.023699 *
team3        0.156366   0.058187   2.687 0.014166 *
team4       -0.139378   0.058187  -2.395 0.026508 *
team5       -0.544639   0.058187  -9.360 9.50e-09 ***
team6       -0.416409   0.058187  -7.156 6.23e-07 ***
saw2        -0.073831   0.058187  -1.269 0.219065
saw3        -0.232082   0.058187  -3.989 0.000723 ***
saw4        -0.052558   0.058187  -0.903 0.377135
saw5         0.033294   0.058187   0.572 0.573573
saw6        -0.146930   0.058187  -2.525 0.020122 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.01015721)

    Null deviance: 4.62947  on 35  degrees of freedom
Residual deviance: 0.20314  on 20  degrees of freedom
AIC: -50.221

Number of Fisher Scoring iterations: 2

> summary(fit1)

Call:
glm(formula = y ~ species:bark + team + saw, family = gaussian(link = identity))
```

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.15957  -0.06465   0.01093   0.05275   0.11339

Coefficients: (1 not defined because of singularities)
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.55356    0.06719  38.006  < 2e-16 ***
team2            0.14246    0.05819   2.448 0.023699 *
team3            0.15637    0.05819   2.687 0.014166 *
team4           -0.13938    0.05819  -2.395 0.026508 *
team5           -0.54464    0.05819  -9.360 9.50e-09 ***
team6           -0.41641    0.05819  -7.156 6.23e-07 ***
saw2            -0.07383    0.05819  -1.269 0.219065
saw3            -0.23208    0.05819  -3.989 0.000723 ***
saw4            -0.05256    0.05819  -0.903 0.377135
saw5             0.03329    0.05819   0.572 0.573573
saw6            -0.14693    0.05819  -2.525 0.020122 *
species1:bark1 -0.42804    0.05819  -7.356 4.15e-07 ***
species2:bark1 -0.60436    0.05819 -10.387 1.67e-09 ***
species3:bark1 -0.09790    0.05819  -1.683 0.108011
species1:bark2 -0.22366    0.05819  -3.844 0.001013 **
species2:bark2 -0.43161    0.05819  -7.418 3.67e-07 ***
species3:bark2       NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.01015721)

    Null deviance: 4.62947  on 35  degrees of freedom
Residual deviance: 0.20314  on 20  degrees of freedom
AIC: -50.221
```

```
Number of Fisher Scoring iterations: 2

> summary(fit2)

Call:
glm(formula = y ~ species + bark + team + saw, family = gaussian(link = identity))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.15702  -0.06817   0.02109   0.06298   0.13119

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.14853    0.06251  34.369  < 2e-16 ***
species2    -0.19213    0.04093  -4.695 0.000111 ***
species3     0.27690    0.04093   6.766 8.46e-07 ***
bark2        0.15835    0.03342   4.739 9.94e-05 ***
team2        0.14246    0.05788   2.461 0.022150 *
team3        0.15637    0.05788   2.702 0.013029 *
team4       -0.13938    0.05788  -2.408 0.024856 *
team5       -0.54464    0.05788  -9.410 3.61e-09 ***
team6       -0.41641    0.05788  -7.195 3.28e-07 ***
saw2        -0.07383    0.05788  -1.276 0.215388
saw3        -0.23208    0.05788  -4.010 0.000589 ***
saw4        -0.05256    0.05788  -0.908 0.373665
saw5         0.03329    0.05788   0.575 0.570962
saw6        -0.14693    0.05788  -2.539 0.018714 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.01004934)
```

```
    Null deviance: 4.62947  on 35  degrees of freedom
Residual deviance: 0.22109  on 22  degrees of freedom
AIC: -51.175

Number of Fisher Scoring iterations: 2

> y <- data[,1]      # Analysis using untransformed times
> fit0 <- glm(y~row+team+saw, family=Gamma(link=log))
> fit1 <- glm(y~species:bark+team+saw, family=Gamma(link=log))
> fit2 <- glm(y~species+bark+team+saw, family=Gamma(link=log))
> summary(fit0)

Call:
glm(formula = y ~ row + team + saw, family = Gamma(link = log))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.16339  -0.06700   0.01043   0.05123   0.11060

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.126415   0.066290  32.078  < 2e-16 ***
row2        -0.178837   0.057408  -3.115  0.00545 **
row3         0.327836   0.057408   5.711 1.37e-05 ***
row4         0.204395   0.057408   3.560  0.00196 **
row5        -0.002404   0.057408  -0.042  0.96701
row6         0.427429   0.057408   7.445 3.47e-07 ***
team2        0.142779   0.057408   2.487  0.02183 *
team3        0.154695   0.057408   2.695  0.01394 *
team4       -0.136857   0.057408  -2.384  0.02715 *
team5       -0.544224   0.057408  -9.480 7.71e-09 ***
team6       -0.416289   0.057408  -7.251 5.14e-07 ***
```

```
saw2         -0.069846   0.057408  -1.217  0.23790
saw3         -0.229767   0.057408  -4.002  0.00070 ***
saw4         -0.051165   0.057408  -0.891  0.38339
saw5          0.037107   0.057408   0.646  0.52539
saw6         -0.144492   0.057408  -2.517  0.02048 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.009887184)

    Null deviance: 4.49901  on 35  degrees of freedom
Residual deviance: 0.20138  on 20  degrees of freedom
AIC: 96.644

Number of Fisher Scoring iterations: 4

> summary(fit1)

Call:
glm(formula = y ~ species:bark + team + saw, family = Gamma(link = log))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.16339  -0.06700   0.01043   0.05123   0.11060

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.55384    0.06629  38.526  < 2e-16 ***
team2          0.14278    0.05741   2.487  0.02183 *
team3          0.15470    0.05741   2.695  0.01394 *
team4         -0.13686    0.05741  -2.384  0.02715 *
team5         -0.54422    0.05741  -9.480 7.71e-09 ***
```

```
team6           -0.41629    0.05741  -7.251 5.14e-07 ***
saw2            -0.06985    0.05741  -1.217  0.23790
saw3            -0.22977    0.05741  -4.002  0.00070 ***
saw4            -0.05117    0.05741  -0.891  0.38339
saw5             0.03711    0.05741   0.646  0.52539
saw6            -0.14449    0.05741  -2.517  0.02048 *
species1:bark1 -0.42743    0.05741  -7.445 3.47e-07 ***
species2:bark1 -0.60627    0.05741 -10.561 1.25e-09 ***
species3:bark1 -0.09959    0.05741  -1.735  0.09816 .
species1:bark2 -0.22303    0.05741  -3.885  0.00092 ***
species2:bark2 -0.42983    0.05741  -7.487 3.19e-07 ***
species3:bark2       NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.009887184)

    Null deviance: 4.49901  on 35  degrees of freedom
Residual deviance: 0.20138  on 20  degrees of freedom
AIC: 96.644

Number of Fisher Scoring iterations: 4

> summary(fit2)

Call:
glm(formula = y ~ species + bark + team + saw, family = Gamma(link = log))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.15888  -0.07031   0.02015   0.06182   0.12645
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.14989    0.06163  34.885  < 2e-16 ***
species2    -0.19309    0.04035  -4.786 8.86e-05 ***
species3     0.27565    0.04035   6.832 7.29e-07 ***
bark2        0.16012    0.03294   4.861 7.39e-05 ***
team2        0.14188    0.05706   2.487 0.020966 *
team3        0.15385    0.05706   2.696 0.013183 *
team4       -0.13860    0.05706  -2.429 0.023754 *
team5       -0.54390    0.05706  -9.533 2.86e-09 ***
team6       -0.41771    0.05706  -7.321 2.49e-07 ***
saw2        -0.07027    0.05706  -1.232 0.231132
saw3        -0.22991    0.05706  -4.029 0.000561 ***
saw4        -0.05239    0.05706  -0.918 0.368505
saw5         0.03602    0.05706   0.631 0.534324
saw6        -0.14348    0.05706  -2.515 0.019721 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.009766541)

    Null deviance: 4.49901  on 35  degrees of freedom
Residual deviance: 0.21901  on 22  degrees of freedom
AIC: 95.668

Number of Fisher Scoring iterations: 4
```