Generalized Linear Models in R

Stats 306a, Winter 2005, Gill Ward

General Setup

- Observe $\mathbf{Y}$ ($n \times 1$) and $\mathbf{X}$ ($n \times p$).
- Assume $\mathbf{Y}$ has an exponential family distribution with some parameterization $\boldsymbol{\zeta}$ known as the **linear predictor**, such that $\boldsymbol{\zeta} = \mathbf{X}\beta$.
- We wish to estimate the parameters $\boldsymbol{\beta}$ ($p \times 1$).
- Let $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$ denote the natural and mean parameterizations of the exponential family.
- There is often a **scale parameter** ϕ.
- The **link function** l is defined by $l(\mu_i) = \zeta_i$.
- The **canonical link** is the function l such that $l(\mu_i) = \eta_i$.

R commands

The R function for fitting a generalized linear model is `glm()`, which is very similar to `lm()`, but which also has a `family` argument. For example:

```
glm( numAcc~roadType+weekDay, family=poisson(link=log), data=roadData)
```

fits a model $Y_i \sim \text{Poisson}(\mu_i)$, where $\log(\mu_i) = \mathbf{X}_i\boldsymbol{\beta}$. Omitting the `link` argument, and setting `family=poisson`, we get the same answer because the log link is the canonical link for the Poisson family. Other families available include `gaussian`, `binomial`, `inverse.gaussian` and `Gamma`. The models are fit using iterative reweighted least squares, so it also possible to set convergence parameters. It is also possible to include an offset term in the formula, using the `offset()` argument in the formula.

As with `lm()`, there are a number of methods for `glm` objects, including `summary`, `coef`, `resid`, `predict`, `anova` and `deviance`. To find out more about these methods type e.g. `help(predict.glm)`.

- `resid`. There are four types of residual, the default of which is `type="deviance"`.
- `predict`. By default the predicted values are of the linear predictor ζ (`type="link"`). Of more use is `type="response"`, and you can also specify `se.fit=T` for the standard errors.

Examples

Binomial. $Y_i \sim \text{Binomial}(n_i, p_i)$, where n_i fixed and $l(p_i) = \mathbf{X}_i\boldsymbol{\beta}$.

There are three ways to specify the number of trials n_i:

- The response is a vector: it is assumed to be of the form y_i/n_i and if the n_i are in the vector `numTrials`, you must also specify `weights=numTrials`.
- The response is a logical vector or factor: it is treated as a binary outcome.
- The response is a two-column matrix: the first column is assumed to the be the number of successes and the second column is the number of failures.

There are three link functions: `logit`, `probit` and `cloglog`.

Example 1: Toxicity of a pyrethroid to the tobacco budworm. 6 levels of doses are given to groups of 20 male and female moths.

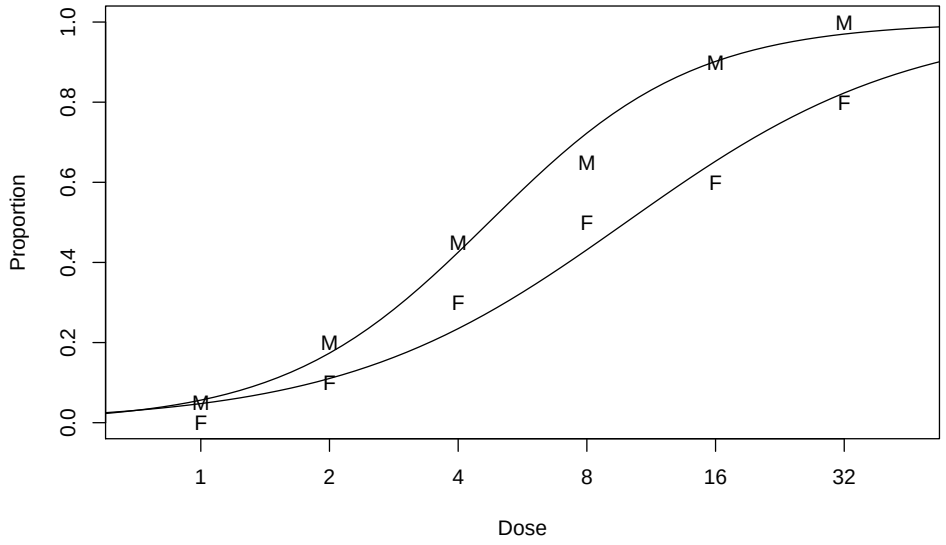

Figure 1: The observed proportion of deaths and predicted probability of death for male and female budworms by dose.

```
> ldose <- rep(0:5,2)
> numdead <- c(1,4,9,13,18,20,0,2,6,10,12,16)
> sex <- factor(rep(c("M","F"),rep(6,2)))
> SF <- cbind(numdead, numalive=20-numdead)
> budworm <- glm( SF ~ sex * ldose, family=binomial)
> summary(budworm)

  Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.9935     0.5527  -5.416 6.09e-08 ***
sexM          0.1750     0.7783   0.225    0.822
ldose         0.9060     0.1671   5.422 5.89e-08 ***
sexM:ldose    0.3529     0.2700   1.307    0.191
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 124.8756  on 11  degrees of freedom
Residual deviance:   4.9937  on  8  degrees of freedom
AIC: 43.104

Number of Fisher Scoring iterations: 4

> anova(budworm, test="Chi")
Analysis of Deviance Table
Terms added sequentially (first to last)

          Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                         11    124.876
sex        1    6.077        10    118.799     0.014
ldose      1  112.042         9      6.757 3.499e-26
sex:ldose  1    1.763         8      4.994     0.184

> summary( glm( SF ~ sex + ldose, family=binomial) )

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.4732     0.4685  -7.413 1.23e-13 ***
sexM          1.1007     0.3558   3.093  0.00198 **
ldose         1.0642     0.1311   8.119 4.70e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 124.876  on 11  degrees of freedom
Residual deviance:   6.757  on  9  degrees of freedom
AIC: 42.867

Number of Fisher Scoring iterations: 4
```

Poisson. $Y_i \sim \text{Poisson}(\lambda_i)$, where $l(\lambda_i) = \mathbf{X}_i\boldsymbol{\beta}$.

There are three link functions: `log`, `identity` and `sqrt`.

Example 2: Death penalty verdicts for cases involving multiple murders in Florida between 1976 and 1987.

		Victim's Race					
		White		Black		Overall	
		Defendant's Race		Defendant's Race		Defendant's Race	
		White	Black	White	Black	White	Black
Death	Yes	53	11	0	4	53	15
Penalty	No	414	37	16	139	430	176
Percent Yes		11.3	22.9	0.0	2.8	11.0	7.9

```
> deathpenalty <- data.frame(
+  number = c(53,11,0,4,414,37,16,139),
+  victim = c("W","W","B","B","W","W","B","B"),
+  defendant = c("W","B","W","B","W","B","W","B"),
+  death = rep(c("yes","no"),rep(4,2)))
>
> summary(glm(number~(victim+defendant+death)^2, family=poisson, data=deathpenalty))

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         4.93578    0.08471  58.265  < 2e-16 ***
victimW            -1.32980    0.18479  -7.196 6.19e-13 ***
defendantW         -2.17465    0.26377  -8.245  < 2e-16 ***
deathyes           -3.59610    0.50691  -7.094 1.30e-12 ***
victimW:defendantW  4.59497    0.31353  14.656  < 2e-16 ***
victimW:deathyes    2.40444    0.60061   4.003 6.25e-05 ***
defendantW:deathyes -0.86780    0.36707  -2.364   0.0181 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1225.07955  on 7  degrees of freedom
Residual deviance:    0.37984  on 1  degrees of freedom
AIC: 52.42

Number of Fisher Scoring iterations: 3
```

```
> deathpenalty2 <- data.frame(
+  prop = c(53,11,0,4)/(c(53,11,0,4)+c(414,37,16,139)),
+  victim = c("W","W","B","B"),
+  defendant = c("W","B","W","B"),
+  weights = c(53,11,0,4)+c(414,37,16,139))
> summary (glm(prop~victim+defendant, family=binomial, weights=weights, data=deathpenalty


Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.5961     0.5069  -7.094 1.30e-12 ***
victimW       2.4044     0.6006   4.003 6.25e-05 ***
defendantW   -0.8678     0.3671  -2.364   0.0181 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22.26591  on 3  degrees of freedom
Residual deviance:  0.37984  on 1  degrees of freedom
AIC: 19.3

Number of Fisher Scoring iterations: 4
```

Example 3: Father's and son's occupational status (UK).

		Son's Status				
		1	2	3	4	5
Father's Status	1	50	45	8	18	8
	2	28	174	84	154	55
	3	11	78	110	223	96
	4	14	150	185	714	447
	5	3	42	72	320	411

```
> mobility <- data.frame(
+  number = c(50,45,8,18,8,28,174,84,154,55,11,78,110,
+  223,96,14,150,185,714,447,3,42,72,320,411),
+  father = factor(rep(1:5,rep(5,5))),
+  son = factor(rep(1:5,5)),
+  diff = factor(abs(rep(1:5,rep(5,5))-rep(1:5,5))),
+  up = factor(rep(1:5,rep(5,5))>rep(1:5,5)))
>
> mobility.glm <- glm(number~father+son+diff+up, family=poisson, data=mobility)
```

```
> anova(mobility.glm,test="Ch")
Analysis of Deviance Table
Terms added sequentially (first to last)

       Df Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                      24     4008.8
father  4   1555.7        20     2453.1        0.0
son     4   1660.9        16      792.2        0.0
diff    4    738.2        12       54.0 1.904e-158
up      1      1.9        11       52.1        0.2


> tapply(round(resid(mobility.glm,type="dev"),1),
+  list(mobility$father,mobility$son), function(x) x)
     1    2    3    4    5
1  4.4 -1.0 -3.4 -1.2  0.6
2 -1.7  0.1 -0.9  0.8  1.0
3 -1.9 -1.2  0.5  1.2 -0.3
4 -1.3  1.1  1.1 -0.9  0.1
5 -0.8  0.6  0.3  0.2 -0.4

> summary(mobility.glm)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.22256    0.12839  25.100  < 2e-16 ***
father2      0.89442    0.10700   8.359  < 2e-16 ***
father3      0.71268    0.11960   5.959 2.54e-09 ***
father4      1.67743    0.13542  12.386  < 2e-16 ***
father5      1.30814    0.17416   7.511 5.86e-14 ***
son2         1.03313    0.11372   9.085  < 2e-16 ***
son3         0.71629    0.12525   5.719 1.07e-08 ***
son4         1.70334    0.13969  12.193  < 2e-16 ***
son5         1.50757    0.17567   8.582  < 2e-16 ***
diff1       -0.47484    0.07252  -6.548 5.85e-11 ***
diff2       -1.01042    0.07969 -12.680  < 2e-16 ***
diff3       -1.92483    0.10990 -17.514  < 2e-16 ***
diff4       -3.02426    0.31289  -9.666  < 2e-16 ***
upTRUE       0.16429    0.11853   1.386    0.166
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4008.809  on 24  degrees of freedom
Residual deviance:   52.115  on 11  degrees of freedom
AIC: 230.84

Number of Fisher Scoring iterations: 4
```

Overdispersion.

- Estimate the overdispersion parameter ϕ directly, where $g(\mathbf{Y};\boldsymbol{\eta},\phi) = \exp\{\,(\boldsymbol{\eta}'\mathbf{Y} - \psi(\boldsymbol{\eta}))/\phi\,\}\, g_0(\mathbf{Y})$.
  Can use `family=quasibinomial` or `family=quasipoisson`.
- Use a negative binomial family where $Y|E \sim$Poisson(μE) and $E \sim$gamma$(\theta)/\theta$. (Also beta binomial family.)

Example 4: The number of days absent from school in a year by children from a large town in rural NSW, Australia. Children were classified by age (4 levels), ethnicity (aboriginal or not), whether they were a slow or fast learner, and sex (M or F).

```
> library(MASS)
> attach(quine)
> round(tapply(Days,list(Lrn,Age,Sex, Eth),var)/
+ tapply(Days,list(Lrn,Age,Sex, Eth),mean),2)
, , F, A

      F0    F1    F2    F3
AL 14.76  3.75    NA 15.15
SL    NA 15.45 19.31    NA

, , M, A

     F0   F1    F2   F3
AL 4.96 2.33  7.87 3.96
SL 4.33 3.00 14.79   NA

, , F, N

     F0   F1   F2   F3
AL 6.14 7.27   NA 9.78
SL   NA 2.90 3.97   NA

, , M, N

      F0   F1   F2    F3
AL  5.53 0.14 9.82 19.27
SL 35.23 5.99 1.68    NA


> quine.nb <- glm.nb(Days~.^4, data=quine)
> quine.nb2 <- stepAIC(quine.nb)
```

```
> summary(quine.nb2,cor=F)
Coefficients: (3 not defined because of singularities)
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)        3.1693     0.3411   9.292  < 2e-16 ***
EthN              -0.3560     0.4210  -0.845 0.397848
SexM              -0.6920     0.4138  -1.672 0.094459 .
AgeF1             -0.6405     0.4638  -1.381 0.167329
AgeF2             -2.4576     0.8675  -2.833 0.004612 **
AgeF3             -0.5880     0.3973  -1.480 0.138885
LrnSL             -1.0264     0.7378  -1.391 0.164179
EthN:SexM         -0.3562     0.3854  -0.924 0.355364
EthN:AgeF1         0.1500     0.5644   0.266 0.790400
EthN:AgeF2        -0.3833     0.5640  -0.680 0.496746
EthN:AgeF3         0.4719     0.4542   1.039 0.298824
EthN:LrnSL         0.9651     0.7753   1.245 0.213255
SexM:AgeF1         0.2985     0.6047   0.494 0.621597
SexM:AgeF2         3.2904     0.8941   3.680 0.000233 ***
SexM:AgeF3         1.5412     0.4548   3.389 0.000702 ***
SexM:LrnSL         0.5457     0.8013   0.681 0.495873
AgeF1:LrnSL        1.6231     0.8222   1.974 0.048373 *
AgeF2:LrnSL        3.8321     1.1054   3.467 0.000527 ***
AgeF3:LrnSL            NA         NA      NA       NA
EthN:SexM:LrnSL    1.3578     0.5914   2.296 0.021684 *
EthN:AgeF1:LrnSL  -2.1013     0.8728  -2.408 0.016058 *
EthN:AgeF2:LrnSL  -1.8260     0.8774  -2.081 0.037426 *
EthN:AgeF3:LrnSL       NA         NA      NA       NA
SexM:AgeF1:LrnSL  -1.1086     0.9409  -1.178 0.238671
SexM:AgeF2:LrnSL  -2.8800     1.1550  -2.493 0.012651 *
SexM:AgeF3:LrnSL       NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.8653) family taken to be 1)

    Null deviance: 265.27  on 145  degrees of freedom
Residual deviance: 167.44  on 123  degrees of freedom
AIC: 1091.4

Number of Fisher Scoring iterations: 1


              Theta:  1.865
          Std. Err.:  0.258
```