Generalized Linear Models

Let Y be a response variable, and $X_1, \cdots, X_p$, predictor variables.

Suppose there is a function g such that

$$g(\mu) = \beta_o + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$Var(Y \mid \mathbf{x}) = \phi \nu(\mu)$$

The function g is known as a *link function*.

ϕ is a dispersion parameter.

Distribution	Family	Link	Variance
Normal/Gaussian	`gaussian`	μ	1
Binomial	`binomial`	$\log(\mu/(1-\mu))$	$\mu(1-\mu)/n$
Poisson	`poisson`	$\log(\mu)$	μ
Gamma	`gamma`	$1/\mu$	μ^2
Inverse Normal/ Gaussian	`inverse.gaussian`	$1/\mu^2$	μ^3
Quasi	`quasi`	$g(\mu)$	$V(\mu)$

Logistic Regression

Let p_x denote the probability of cancer given the smoking status of the individual. Then

$$p_x = \frac{e^{\beta_0+\beta_1 X_1}}{1+e^{\beta_0+\beta_1 X_1}}$$

In terms of the link function

$$logit(p_x) = \beta_0 + \beta_1 x$$

Denote $$X_1 = \begin{cases} 1, & \text{Smoker} \\ 0, & \text{Non-smoker} \end{cases}$$

β_1 is the *log odds ratio* of having cancer for a smoker relative to a non-smoker.

Likelihood Inference

Suppose $y \sim f(y;\theta)$. The log likelihood is given by

$$L(\theta) = \sum_{j=1}^{n} \ln f(y_j;\theta)$$

The *deviance* is defined as

$$D(y;\theta) = 2\phi[L(y) - L(\theta)]$$

where $L(y)$ is the saturated model.

In the Gaussian case, when $\hat{\theta}$ is the m.l.e., the deviance corresponds to the RSS.

Analysis of Deviance

Sums of squares for non-normal data are not appropriate measures of contributions of a sum to total variation.

Suppose $\boldsymbol{\theta_1}$ and θ_2 correspond to two competing models.

Difference in deviance is given by

$$D(\boldsymbol{\theta_1}; \theta_2) = D(y; \boldsymbol{\theta_1}) - D(y; \theta_2)$$

Under $\boldsymbol{\theta_1}$, $D(\boldsymbol{\theta_1}; \theta_2)$ is approximately χ^2_ν, where $\nu = \boldsymbol{\nu_1} - \nu_2$, the difference in the corresponding model degrees of freedom.

For model selection, one would reject the $\boldsymbol{\theta_1}$ model, if the difference is too large, i.e., the model based on θ_2 fits better.

Residuals

- Response residuals

$$r_j^R = (y_j - \hat{\mu}_j)$$

- Pearson residuals

$$r_j^P = \frac{(y_j - \hat{\mu}_j)}{\sqrt{Var(\hat{\mu}_j)}}$$

- Deviance residuals

  Let d_j denote the contribution of the j'th observation to the deviance. Then

$$r_j^D = sgn(y_j - \hat{\mu}_j)\sqrt{d_j}$$

  indicates the influence of the j'th observation to the fit.

Model Selection in GLM

S-PLUS:

The step.glm() and step() functions allow model selection in glm.

R functions:

```
> library(help="MASS")
> help(stepAIC)
```

Poisson regression

$$\ln(\mu) = \beta_o + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Most appropriate for rare events, e.g., when $\lambda = 0$, 1 or 2.

  When λ is large, histogram more symmetric.

  Lognormal, gamma or $\sqrt{Y}$.

- Poisson Regression: Model number of occurrences of an event or rate as a function of the predictors
  - Number of ear infections in infants in a given period.
  - Rate of insurance claims
  - Homicide rate
  - Number of equipment failures in a given period.

Interpretation of Coefficients

Consider the Poisson model:

$$\ln(\mu) = \beta_o + \beta_1 X_1$$

Then e^{β_1} is the multiplicative effect on μ corresponding to a unit change in X_1.

When X_1 takes the values 1 and 0

Suppose $e^{\hat{\beta}_1} = 0.85$, and Y is the number of seizure attacks per week.

Treatment reduces the mean number of attacks by 15%.

Overdispersion

Sample mean substantially smaller than variance.

- Overdispersion leads to underestimation of SE (inflation of Type I error rate)
- Caused by: Model under-specification (some relevant predictors not in model)
  Outliers
  Clustered data (e.g., colony of bacteria)

Overdispersion (cont)

- Measures:
  Adjusted SE

$$SE(\hat{\beta})_{adj} = \sqrt{\hat{\phi}} SE(\hat{\beta})_{unadj}$$

  where $\hat{\phi} =$ Deviance (Pearson) χ^2/df
  Remove outliers
  Check mis-specification
  Use negative binomial distribution
- MLE for $\beta's$ still correct

Negative Binomial Distribution

Appropriate for aggregate events.

Example: Distribution of species, people, animals in space

Probability mass function

$$f(k) = \frac{\Gamma(r+k)}{\Gamma(k+1)\Gamma(r)} p^r (1-p)^k$$

where $r > 0$, and $0 < p < 1$ are parameters.

Mean:

$$\mu = r\frac{1-p}{p}$$

Variance

$$\sigma^2 = r\frac{1-p}{p^2}$$

Clearly $\sigma^2 > \mu$.

Special Cases

- Pascal Distribution: In $k + r$ Bernoulli trials, let k = number of failures before the the rth success.
- Geometric Distribution: The probability of k failures before the first success.

Alternatively, if λ has a Gamma distribution with parameters r and $(1-p)/p$;

and $f(k \mid \lambda)$ is Poisson with mean λ.

Then f(k) is negative binomial.

Implementation

- R

glm(formula, family=poisson, data=DATASET, offset … contrasts=NULL, ...)

fit1 <- glm(skips ~ ., family = poisson, data = solder.balance)

anova(*fit1*, test = "Chi")

Implementation

- SAS

```
proc genmod data = DATASET;
 class VAR1 VAR2;
 model Y = VAR1 VAR2 VAR3/ dist = poisson link = log;
 estimate 'LABEL' VAR1 1 -1/ exp;
run;
```

If data is given as rates:

$$\ln\left(\frac{\mu}{n}\right) = \beta_0 + \beta_1 X_1$$

$$\ln(\mu) = \beta_0 + \beta_1 X_1 + \ln(n)$$

log(y) = x'b + log(offset)

proc genmod data = DATASET;
class VAR1 VAR2 ID;
model Y = VAR1/ dist = poisson
link = log
offset=ln
type3;

In R:

```
anorex.1 <- glm(Y ~ formula + offset(log(offset_var)),
                family = poisson, data = dat)
```

Logit Analysis of Longitudinal (Panel) and other Clustered Data

Example: Subjects randomized to one of two depression drugs. Response recorded as *Improved* or *Not Improved"* at Months 3, 6 and 9 following initial treatment.

Standard logistic regression analysis not appropriate.

- SE underestimated
- Coefficient estimates inefficient.
  There exist estimates with lower SE's.

Approaches

- GEE methods
  Need to specify link function and "working" correlation matrix.
  Examples of "working" correlation matrix: *Independence, Exchangeable, Unstructured, AR(1), etc.*
  Method is robust against misspecification of correlation matrix.

```
proc genmod data = DATASET;
      class VAR1;
      model Y = VAR1/ dist = poisson
                       link = log
                      offset=ln
                       type3;
      repeated subject = id/ type = unstr;
      estimate ‘LABEL' VAR1 1 -1/ exp; run;
```

R Commands:

```
library(geepack)
help(package=“geepack”)
help(geeglm)
```

	y	trt	base	age	V4	subject	period	lbase	lage	time
1	5	placebo	11	31	0	1	1	-0.7563538	0.11420370	1
2	3	placebo	11	31	0	1	2	-0.7563538	0.11420370	1
3	3	placebo	11	31	0	1	3	-0.7563538	0.11420370	1
4	3	placebo	11	31	1	1	4	-0.7563538	0.11420370	1
5	3	placebo	11	30	0	2	1	-0.7563538	0.08141387	1

```
library(MASS)
attach(epil)
#Consider Period 4 Data
> Y4 <-y[V4==1]
> TRT <-trt[V4==1]
> LOGAGE <-lage[V4==1]
> LOGBASE <-lbase[V4==1]
```

- `TRT <-1*(TRT=="progabide")`

```
> summary(glm(Y4~TRT,family="poisson"))
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.07497    0.06696  30.986   <2e-16 ***
TRT         -0.17142    0.09640  -1.778   0.0754 .
```

```
summary(glm(Y4~LOGBASE*TRT,family="poisson"))
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.83844    0.08251  22.281    < 2e-16 ***
LOGBASE      0.93454    0.08991  10.394  < 2e-16 ***
TRT         -0.48013    0.12938  -3.711    0.000206 ***
LOGBASE:TRT  0.43    0.13   3.3          0.000860 ***
```

SAS

```
data epilepsy;input
id     y    trt $ base age V4 subject period   lbase    lage time;
cards;
1     5   placebo   11  31  0       1  1 -0.75635379  0.11420370    1
2     3   placebo   11  31  0       1  2 -0.75635379  0.11420370    1
3     3   placebo   11  31  0       1  3 -0.75635379  0.11420370    1
......
;
run;
data epi4;set epilepsy;
     if V4=1;

proc genmod data= epi4;
          class trt;
 model y  = trt / dist=poisson;
run;
```

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	1.9036	0.0693	1.7677	2.0394	753.69	<.0001
trt	placebo	1	0.1714	0.0964	-0.0175	0.3603	3.16	0.0754
trt	progabid	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

```
proc genmod data=epilepsy ;
      class trt ; model y= trt / dist=poisson; repeated subject=subject /
 type=exch ; run;
```

GEE Model Information

Correlation Structure	Exchangeable
Subject Effect	subject (59 levels)
Number of Clusters	59
Correlation Matrix Dimension	4
Maximum Cluster Size	4
Minimum Cluster Size	4

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		2.0744	0.2990	1.4883	2.6604	6.94	<.0001
trt	placebo	0.0751	0.3539	-0.6185	0.7687	0.21	0.8320
trt	progabid	0.0000	0.0000	0.0000	0.0000	.	.

Cumulative Logit Model

Example. Suppose wish to study the association between Y (the response variable denoting the state of depression: *None*, *Mild*, *Severe*) and the explanatory variable *Gender*.

Denote the categories of Y by j (with j=1, None; 2 Mild; and 3 Severe).

Let p_j denote the probability of falling into category j of Y.

Define the *cumulative probabilities*

$$F_j = \sum_{m=1}^{j} p_j, \quad j = 1, \cdots J-1$$

i.e. the probability of falling in jth category or lower.

The *cumulative logit model*

$$\ln(\frac{F_j}{1-F_j}) = \alpha_j + \beta_1 X_1, \quad j = 1, \cdots J-1$$

- Interpretation of α_j not relevant.
- As J increases, too many α_j's to estimate. Rule-of-thumb: 10 observations per parameter
- *Proportional Odds Assumptions*

  Need whether the ordinal restriction is valid.

  Test based on fitting different logistic models, with Y dichotomized differently.

  Test whether the coefficients corresponding to the different models are equal.

Implementation in SAS

```
proc genmod data = DATASET;
    class VAR1;
    model Y = VAR1/ dist = multinomial
                    link = cumlogit
                      type3;
```

- Multinomial Logit Analysis
  - Used when the J categories are not ordinal
  - PROC CATMOD in SAS

Probit Regression

An alternative to logistic regression.

Example. In toxicity studies, let p_x denote the probability of death when the dose $X = x$.

Then $p_x = \Phi(\beta_o + \beta_1 x)$

The link function:

$$\Phi^{-1}(p_x) = \beta_o + \beta_1 x$$

To estimate the median of the tolerance distribution, LD_{50}:

$$\Phi^{-1}(0.50) = \beta_o + \beta_1 x$$

gives $\hat{x}_{50} = -\hat{\beta}_o / \hat{\beta}_1$.

Suppose we use logit link:

$$\ln\left(\frac{0.50}{1-0.50}\right) = \beta_o + \beta_1 x$$

also gives $\hat{x}_{50} = -\hat{\beta}_o/\hat{\beta}_1$.

NB: Logistic and probit models often similar. When response is concentrated in the tail, the two different.

glm(formula, family=binomial(link=probit),...)

Problem Set 9

1. Reading Assignment: Ramsey and Schafer, Chapter 22, pp. 644-668
2. Ramsey and Schafer. Problem Number 25, Page 667 (Body Size and Reproductive Success).