

Session 7

Generalised Linear Models

	<i>page</i>
Examples of GLMs in Medical Statistics	7-4
The GLM Algorithm	7-5
Specifications in STATA	7-8
Main Output from STATA	7-10
Example-Coronary Heart Disease Data	7-11

Generalised Linear Models

Three components:

- 1.A **probability distribution** D for the y_i
 D is from the exponential family

$$E(y_i) = \mu_i$$

- 2.A **linear predictor** η_i

$$\eta_i = \sum \beta_j x_{ij}$$

- 3.A **link function** $g_i(\cdot)$

$$g_i(\mu_i) = \eta_i$$

usually

g_i is known

g_i is same for
all

observations

||

Choice of distribution D includes

Normal	}	continuous data
Exponential		
Gamma		
Inverse Gaussian		
Poisson	-	count data
Bernoulli	}	binary data (yes/no)
Binomial		

D may have a scale parameter ϕ

Choice of link function $g(\cdot)$ includes:

Identity	$\mu_i = \eta_i$
Log	$\log(\mu_i) = \eta_i$
logit	$\log\left(\frac{\mu_i}{1 - \mu_i}\right) = \eta_i$

Examples of GLMs in Medical Statistics

Logistic Regression

Distribution Binomial or Bernoulli

Link Logit

Response $0 \dots N_i$ or 0,1

Matched case-control analysis

Conditional logistic regression fitted as GLM

Distribution Poisson

Link Log

Response Case/control (1/0)

Survival Analysis/Event History analysis

Analysis of Person-Epochs

Distribution Poisson

Link Log

Response: (1/0) event occurs within person-epoch(1/0)

The GLM Algorithm

response vector $\underline{y}=[y_i]$ link function $g(\cdot)$
 distribution $D(\cdot)$ model matrix X

$\mu_i = E(y_i)$ $\eta_i = g(\mu_i)$ $\eta = X\beta$
 fitted linear
 values predictor

$$\tau_i^2 = v_i = \frac{\text{var}(y_i)}{\phi} \qquad \frac{\partial \eta_i}{\partial \mu_i} = g'(\mu_i)$$

Then:

$$\left[\sum_i u_i x_{ij} x_{ik} \right] \hat{\beta}_j^{(n+1)} = \left[\sum_i u_i z_i x_{ik} \right]$$

$$(\hat{X}'UX)\hat{\beta} = X'UZ$$

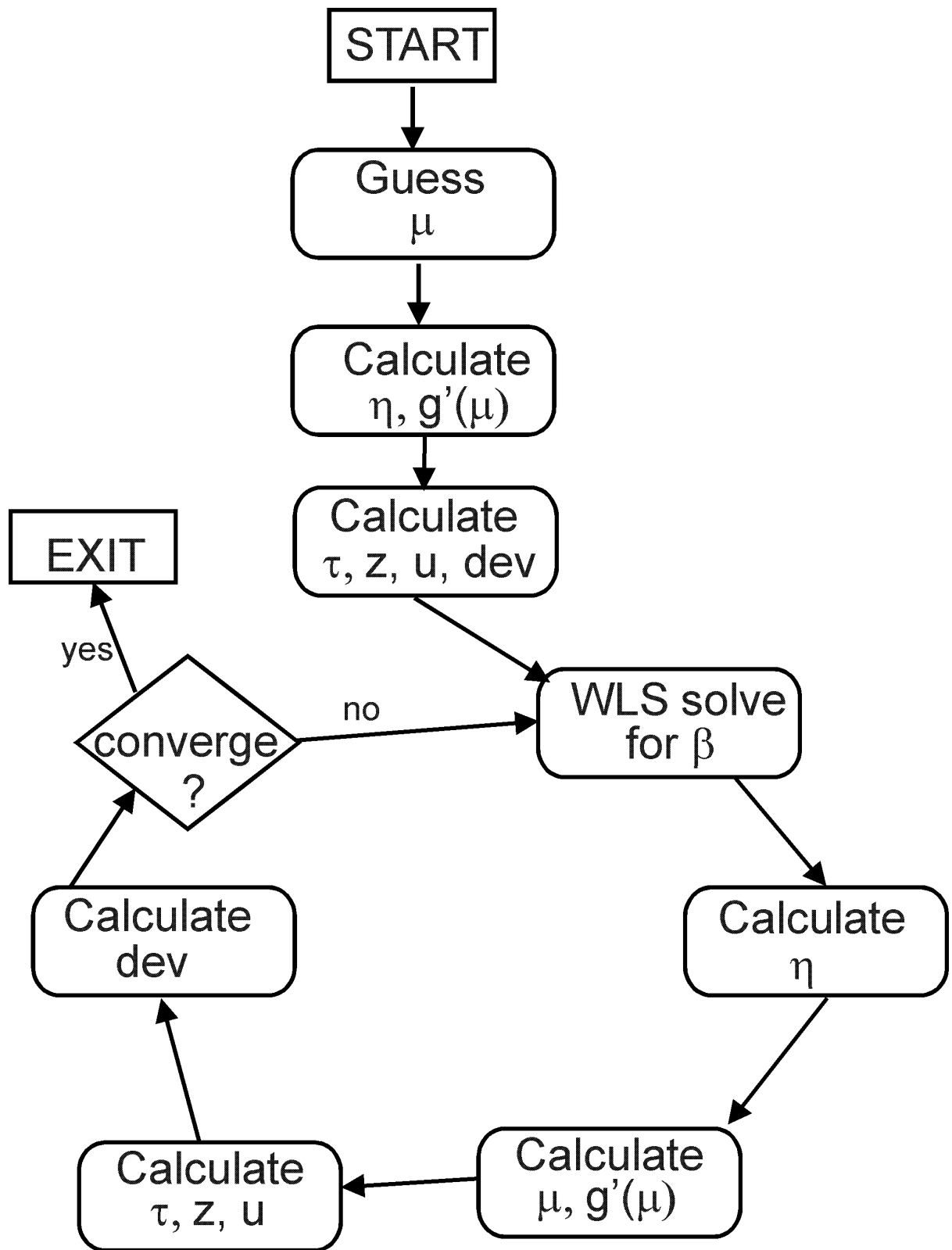
where:

$$u_i = \frac{1}{v_i [g'_i(\mu_i)]^2} \qquad \text{'iterative weights'}$$

$$z_i = \hat{\eta}_i + g'(\mu_i)(y_i - \mu_i) \qquad \text{'working vector'}$$

Weighted least squares algorithm

Weights u_i and adjusted y-variate z_i depend on current fitted values



What is the deviance?

$$\begin{aligned} \text{Scaled deviance} &= -2 \log \left[\frac{L_{\text{model}}}{L_{\text{saturated}}} \right] \\ &= 2 \log L_{\text{saturated}} - 2 \log L_{\text{model}} \end{aligned}$$

model

What is a saturated model?

This is a model with one parameter for every observation. In a saturated model, the fitted values will be equal to the observed y .

A saturated model has a (scaled) deviance of zero.

Specification in STATA

`glm response explanators , options`

response specifies the response variable

explanators specifies a list of explanatory variables,
separated by spaces.

options specify

1. the probability distribution

<code>family(gau)</code>	Normal
<code>family(p)</code>	Poisson
<code>family(b)</code>	Bionomial
<code>family(ig)</code>	Inverse Gaussian
<code>family(gam)</code>	Gamma

2. the link function

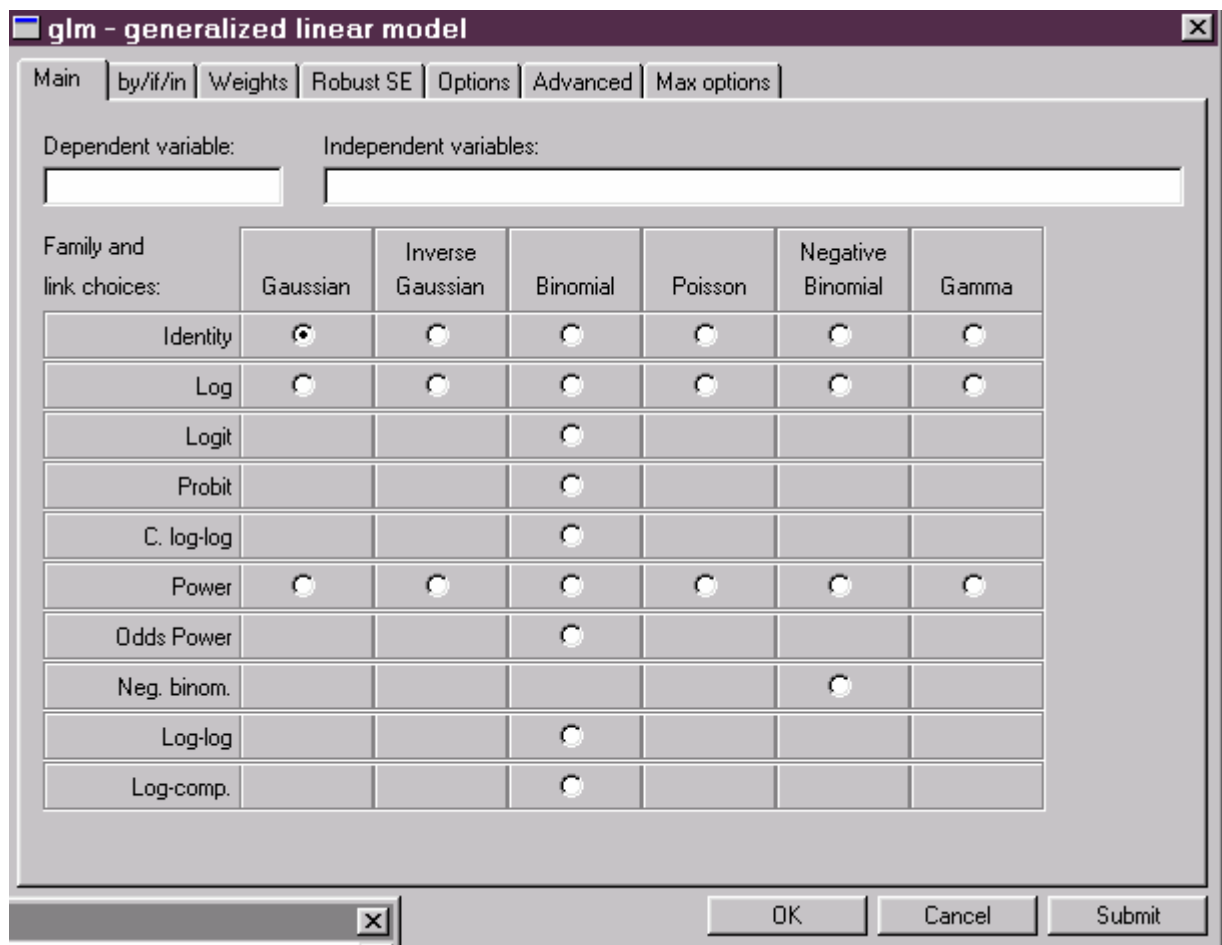
<code>link(identity)</code>	Identity	μ_i
<code>link(log)</code>	log	$\log(\mu_i)$
<code>link(power -1)</code>	reciprocal	$1 / \mu_i$
<code>link(power 0.5)</code>	square root	$\sqrt{\mu_i}$
\vdots	\vdots	

Through the graphical front end, it is slightly easier

statistics>

generalised linear models>

generalised linear models



Note that only certain combinations of distribution and link are allowed.

Main Output from STATA

1. Scaled Deviance or Deviance
if scale parameter if scale
parameter parameter
fixed not fixed

2. Degrees of freedom df
no. of observations in fit - no. of
parameters.

3. Estimates of β s with their standard errors.

4. `predict fv, mu` stores fitted values in `fv`
 $\hat{\mu}_i$

5. `predict res, pearson` stores pearson

6. residuals in `res`

$$\frac{y_i - \hat{\mu}_i}{V(\hat{\mu}_i)}$$

`predict lp, xb` stores linear predictor in `lp`

$$\eta_i = \sum \hat{\beta}_j x_{ij}$$

Or through

Example – Coronary heart disease data

Previously , we used `blogit` command.
Recall – we fit saturated model (all two-way interaction model)

```
xi: blogit r n i.chol i.bp i.chol*i.bp
```

warning message!!

Why does this give the correct likelihood?

For binomial data

$$\log L = \sum_i [r_i \log p_i + (n_i - r_i) \log(1 - p_i)]$$

The contribution of observation i to likelihood is

$$r_i \log p_i + (n_i - r_i) \log(1 - p_i)$$

In a saturated model, $p_i = r_i / n_i$

contribution is $r_i \log r_i + (n_i - r_i) \log(n_i - r_i) - n_i \log(n_i)$

In general, this is not zero, except when $r_i = 0$
or $r_i = n_i$
or $n_i = 1$

So, by omitting an observation with $r_i = 0$ from the fit, the likelihood is still correct, although the df is wrong.

Now we use glm command.

```
xi: glm r i.bp i.chol i.bp*i.chol, family(binomial n ) link(logit)
i.bp          _Ibp_1-4          (naturally coded; _Ibp_1 omitted)
i.chol        _Ichol_1-4        (naturally coded; _Ichol_1 omitted)
i.bp*i.chol   _IbpXcho_#_#      (coded as above)
```

```
Iteration 0:   log likelihood = -25.732689
Iteration 1:   log likelihood = -25.566649
Iteration 2:   log likelihood = -25.555202
Iteration 3:   log likelihood = -25.552663
Iteration 4:   log likelihood = -25.552099
Iteration 5:   log likelihood = -25.551956
Iteration 6:   log likelihood = -25.551929
Iteration 7:   log likelihood = -25.551925
Iteration 8:   log likelihood = -25.551924
```

Generalized linear models	No. of obs	=	16
Optimization : ML: Newton-Raphson	Residual df	=	0
	Scale parameter	=	1
Deviance = 5.71545e-07	(1/df) Deviance	=	.
Pearson = 3.81994e-07	(1/df) Pearson	=	.

Variance function: $V(u) = u*(1-u/n)$	[Binomial]
Link function : $g(u) = \ln(u/(n-u))$	[Logit]
Standard errors : OIM	


```
predict fv,mu
```

```
list fv r
```

	fv	r
1.	3	3
2.	7	7
3.	2	2
4.	8	8
5.	3	3
6.	11	11
7.	2	2
8.	12	12
9.	6	6
10.	11	11
11.	1.48e-07	0
12.	3	3
13.	4	4
14.	3	3
15.	11	11
16.	6	6

glm gives correct results for grouped data -avoid use of **blogit** in STATA