Modern Applied Statistical Models
Topic II: Generalised Linear Models

DT9209: MSc Applied Mathematics

Dr Joe Condon

School of Mathematical Sciences
Dublin Institute of technology

Generalised Linear Models

For linear models we require that the responses are normally distributed. But often this is not the case, from both observational studies and designed experiments.

Example: Space Shuttle O-rings; The data below show the temperature and atmospheric pressure at takeoff for 23 space shuttle flights. Also shown are the number of O-ring failures from the shuttle's six O-rings on the solid rocket boosters. Failure of these O-rings were the cause of the Challenger disaster in 1986.

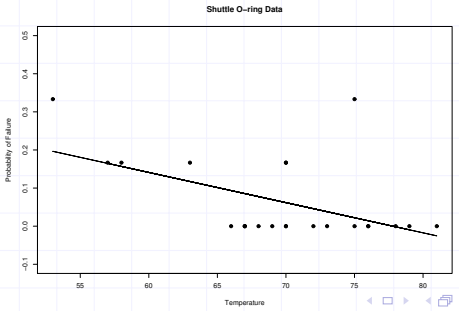
Temp	Pressure	Failures	Temp	Pressure	Failures
66	50	0	67	200	0
70	50	1	53	200	2
69	50	0	67	200	0
68	50	0	75	200	0
67	50	0	70	200	0
72	50	0	81	200	0
73	50	0	76	200	0
70	100	0	79	200	0
57	100	1	75	200	2
63	200	1	58	200	1
70	200	1	76	200	0
78	200	0			

The question, is the O-ring failure related to take-off temperature?

One possibility is to model the probability of a O-ring failure as a function of temperature, i.e. $\pi_i = f(\beta_0 + \beta x_i)$ where the β's are unknown parameters and x is temperature and $\pi_i =$ (Probability of 0-ring failure | temperature x_i).

If we choose $f(.)$ as the identity function we end up with a simple linear regression model.

Applying this to the O-ring data we get the following plot;



There are immediate problems with this linear model;

- The fitted values at high temperature are negative probabilities.
- The probabilities close to zero cannot be normally distributed as negative probabilities are not allowed (same argument for probabilities close to 1).
- If the response are number of successes from a finite number of trials then they are binomial random variables and not normal.

Basically, we have fitted a normality assumption to data that is binomial.

The solution to this modelling problem is to use the **logistic regression model.**

Logistic Regression

We assume that we have the following:

- Independent responses y_i's which follow a binomial (Bernoulli) distribution and represent the number of 'successes' from n_i trials.
- A set of potential predictors that are related to the responses, x_i's.
- The binomial probability π_i changes with the predictors x_i's.

Recall from basic probability that the binomial mass function is;

$$P(Y_i = y_i) \quad = \quad \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i}$$

The mean of the binomial is $n_i \pi_i$ and the variance is $n\pi(1-\pi)$. So, in the above model, if the value of π changes across the observation then the variance cannot be constant - so another linear model assumption is violated.

Logit function

We need to relate the predictors in a regression type way to the mean of the responses. For this, the **logit** link function is used.

$$x_i'\beta = \log\left(\frac{\pi_i}{1-\pi_i}\right) \quad \text{[[the logit of } \pi_i\text{]}$$

$$\Rightarrow \pi_i = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}$$

$$\Rightarrow E[y_i] = n_i \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}$$

Consider the domain and codomain of the logit function...

Fitting the model: Method of Maximum Likelihood

The likelihood is joint probability (or probability density) of the data considered as a function of the unknown parameters:

$$L(\beta; Y, X) \quad = \quad \prod_{i=1}^{n} \binom{n_i}{y_i} \left[\frac{exp(x_i'\beta)}{1+exp(x_i'\beta)}\right]^{y_i} \left[\frac{1}{1+exp(x_i'\beta)}\right]^{n_i - y_i} \quad (1)$$

Since the likelihood is positive and real valued, the (hopefully) unique maximisers of the log of the likelihood will be the same as the likelihood - so use the log likelihood as this is easier to work with.

$$\ell(\beta; y) \quad = \quad \sum_{i=1}^{n} \left\{ \log \binom{n_i}{y_i} + y_i(x_i'\beta) - n_i \log[1+exp(x_i'\beta)] \right\} \quad (2)$$

Where we can evaluate (2) using the data and given values for β. We now need to find those values of β that will maximise (2), however there is no closed form solution. Therefore a general method for maximising non-linear equations would be helpful - one such is the Newton-Rahpson method.

[As an exercise show that the Least Squares solution is the appropriate maximiser of the log likelihood for normally distributed responses.]

Newton-Raphson Method

The log-likelihood in (2) may be viewed as a function of a number of unknown parameter estimates (the β's).

The Newton-Raphson (N-R) method is a general numerical algorithm for iterating to the solution is such cases.

You may have seen the scalar version of N-R: given a function $f(z) = 0$ to find the roots, start with a guess and iterate updating each time using the formula;

$$z^{update} = z^{old} - \frac{f(z^{old})}{f'(z^{old})}$$

For example; What is the value of z such that $exp(-z^2) = z$? Then we have, $f(z) = z - e^{-z^2} = 0$ and $f'(z) = 1 + 2ze^{-z^2}$, so starting at a guess $z = 2$ we iterate to a solution,

z^{old}	$f(z^{old})$	$f'(z^{old})$	z^{new}
2	1.981684361	1.073262556	0.153588466
0.153588466	-0.823098172	1.300015606	0.786733306
0.786733306	0.24822335	1.847327436	0.652364409
0.652364409	-0.001026743	1.852498265	0.652918656
0.652918656	2.97708E-08	1.852605505	0.652918640

Convergence is reached where there is little change in the estimate. So, decide to stop the algorithm where e.g. there is no change correct to 5 decimal places in two successive iterations.

In the case of a GLM $f(z) = 0$ is derived from the log likelihood. The slope of the log likelihood evaluated at the maximum likelihood value of the parameters estimates should be zero - since it is a maximum. So the function we need to use is the first derivative of the log likelihood with respect of the unknown parameters.

These derivatives may be expressed in terms of the β's as a gradient vector (i.e. a vector of partial derivatives),

$$U = \begin{pmatrix} \frac{\partial \ell(\theta;y)}{\partial \beta_0} \\ \frac{\partial \ell(\theta;y)}{\partial \beta_1} \\ \vdots \\ \frac{\partial \ell(\theta;y)}{\partial \beta_k} \end{pmatrix} = 0$$

U plays a large role in GLM and is called the score vector.

$f'(z)$ is the derivative of this score vector - i.e. is the Hessian matrix of second partial derivatives,

$$U' = H = \begin{pmatrix} \frac{\partial^2 \ell(\theta;y)}{\partial \beta_0^2} & \frac{\partial^2 \ell(\theta;y)}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 \ell(\theta;y)}{\partial \beta_0 \partial \beta_2} & \cdots & \frac{\partial^2 \ell(\theta;y)}{\partial \beta_0 \partial \beta_k} \\ \frac{\partial^2 \ell(\theta;y)}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 \ell(\theta;y)}{\partial \beta_1 \partial \beta_1} & \cdots & & \vdots \\ \vdots & & \ddots & & \\ & & & & \frac{\partial^2 \ell(\theta;y)}{\partial \beta_k \partial \beta_k} \end{pmatrix}$$

This leads to the multivariate version of the scalar N-R scheme being applied to GLMs as follows;

Given some prior estimate of the parameters $\hat{\beta}^{(m-1)}$ calculate an update as,

$$\hat{\beta}^{(m)} = \hat{\beta}^{(m-1)} - {H^{(m-1)}}^{-1} U^{(m-1)}$$

Where the score and hessian matrices are evaluated at the $(m-1)$ estimates. Iterate using this scheme until there is little change in the parameter estimates.

This is multivariable version of N-R.

Example: The O-ring data. Lets say we want to fit a regression model to the binomial O-ring data. We will use the log likelihood given above and presume that the linear predictor is a simple linear regression model of the probability of failure on temperature.

The log likelihood for the model is therefore;

$$\ell(\beta; y) = \sum_{i=1}^{n} \left\{ \log \binom{n_i}{y_i} + y_i(\beta_0 + \beta_1 x_i) - n_i \log[1 + exp(\beta_0 + \beta_1 x_i)] \right\}$$

where x_i is the temperature at take-off.

We need to derive expressions for the score vector and the Hessian matrix. Define: $\eta_i = \beta_0 + \beta_1 x_i$ - i.e. the linear predictor.

$$U = \begin{pmatrix} \frac{\partial \ell(\beta;y)}{\partial \beta_0} \\ \\ \frac{\partial \ell(\beta;y)}{\partial \beta_1} \end{pmatrix} = \begin{pmatrix} \sum_i \left\{ y_i - \frac{n_i exp(\eta_i)}{1+exp(\eta_i)} \right\} \\ \\ \sum_i \left\{ (y_i x_i - \frac{n_i x_i exp(\eta_i)}{1+exp(\eta_i)} \right\} \end{pmatrix}$$

and an expression for the hessian can be also be derived,

$$H = \begin{pmatrix} \sum_i \left\{ \frac{n_i exp(\eta_i)^2}{[1+exp(\eta_i)]^2} - \frac{n_i exp(\eta_i)}{1+exp(\eta_i)} \right\} & \sum_i \left\{ \frac{n_i x_i exp(\eta_i)^2}{[1+exp(\eta_i)]^2} - \frac{n_i x_i exp(\eta_i)}{1+exp(\eta_i)} \right\} \\ \\ \sum_i \left\{ \frac{n_i x_i exp(\eta_i)^2}{[1+exp(\eta_i)]^2} - \frac{n_i x_i exp(\eta_i)}{1+exp(\eta_i)} \right\} & \sum_i \left\{ \frac{n_i x_i^2 exp(\eta_i)^2}{[1+exp(\eta_i)]^2} - \frac{n_i x_i^2 exp(\eta_i)}{1+exp(\eta_i)} \right\} \end{pmatrix}$$

When can now apply the method to the O-ring data starting at reasonable starting values for the parameter estimates - $\beta_0 = 1, \beta_1 = 0$,
Iteration 1

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} -27.13 & -1887.48 \\ -1887.47 & -132595.09 \end{pmatrix}^{-1} \begin{pmatrix} -91.89 \\ -6444.16 \end{pmatrix} = \begin{pmatrix} 0.4168 \\ -0.0403 \end{pmatrix}$$

Iteration 2 etc...

and convergence is reached to 4 decimal places after 5 iterations.

NB: The N-R method is quite a fast algorithm. It works well with GLM problems as they tend to be 'well' behaved (i.e. satisfy some technical conditions).

It may fail where data is sparse - e.g. nearly all zero responses. The other main method which can be equivalent but has a different development is Iterative Weighted Least Squares (IWLS). SAS uses a ridge-stabilised version of the N-R algorithm - in many cases it really amounts to the same thing.

Interpretation of Parameters from Logistic Regression Model

What do the parameters from a logistic regression model represent?

We have had;

$$g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i'\beta = \eta_i$$

So, the linear predictor from a logistic regression model (i.e $x_i'\beta$) is the logit of π_i for a given x_i'.

The logit is also the log odds for a binary success/failure situation. Recall some definitions in the binary case;

1. Probability: $\frac{\text{No of ways to success}}{\text{No of ways to success + No of ways to failure}}$, e.g. $p = .66^{\cdot}$ means in the long run $\frac{2}{3}$ experiments will result in a successes. Range = (0,1)

2. Odds: $\frac{\text{No of ways to success}}{\text{No of ways to failure}}$, e.g. odds = 2 means success is twice as likely as failure(same as probability = $0.66^{\cdot}$). Range=(0,∞).

3. Odds ratio = $\frac{\text{Odds in experiment 1}}{\text{Odds in experiment 2}}$, e.g. odds ratio = 0.5 means the odds of a success in experiment 1 is half that of experiment 2. Range=(0,∞).

Odds ratio are very popular in health information, as they tend to accentuate difference - e.g. - fictitious data -

	Cancer	No Cancer	
Smoker	1,000	2,000	3,000
Non Smoker	1,000	5,000	6,000

So, twice as likely to develop cancer if you smoke - or the odds of a smoker getting cancer is 2.5 times the non smoker.

The parameter estimates therefore can be used to derive the log odds of success given a set of covariate values.

E.g. In the O-ring data, the estimated log odds of a particular O-ring failing at a take-off temperature of 60 degrees is;

$$\begin{pmatrix} 1 & 60 \end{pmatrix} \begin{pmatrix} 5.0850 \\ -0.1156 \end{pmatrix} \quad = \quad -1.851$$

So the odds of the O-ring failing is $e^{-1.851} = 0.16$.

NB. here were are defining the binomial success as an O-ring failure.

There is a more intuitive interpretation of the slope parameter.

Imagine there are two launches of the shuttle, launch A has a takeoff temperature of 61 degrees and launch B has a takeoff temperature of 60 degrees. Then we have the following relationship for the log odds of the probability of failure on two flights:

$$\log \left[\frac{\pi_a}{1-\pi_a} \Big/ \frac{\pi_b}{1-\pi_b} \right] = \beta_1$$

From this we get that the parameter estimate $\hat{\beta}$ is an estimate of the population log odds-ratio for a unit increase in x.

NB. it does not matter what the actual two values of x are, as long as they are 1 unit apart.

E.g. What is the estimated change in the odds ratio between a take-off at 60 degrees and 80 degrees?

The log odds ratio is given by; $\hat{\beta}_1 \times -20 \Rightarrow -0.1156 \times -20 = 2.312$.

So the odds ratio is $e^{2.312} = 10.09$

- i.e. the odds of the O-ring failing at the lower take-off temperature is ten times that of the higher temperature.

We can also use the parameters to calculate probabilities if you prefer (statisticians tend to prefer probabilities).

Given a set of covariates x_i' simply solve the log odds equation for $\hat{\pi}$,

$$log \frac{\hat{\pi}_i}{1-\hat{\pi}_i} = x_i' \hat{\beta} = \hat{\eta}_i$$

$$\Rightarrow \hat{\pi}_i = \frac{e^{\hat{\eta}_i}}{1+e^{\hat{\eta}_i}} \qquad (3)$$

e.g. what is the estimated probability for a given O-ring failing with a take-off temp. of 72 degrees?

what is the estimated probability for a given O-ring failing with a takeoff temp. of 54 degrees?

Fitting the model using R

There are a number of functions in procedures in R for fitting GLMs. One of the more versatile is glm() function. The syntax for glm() is not unlike that of lm():

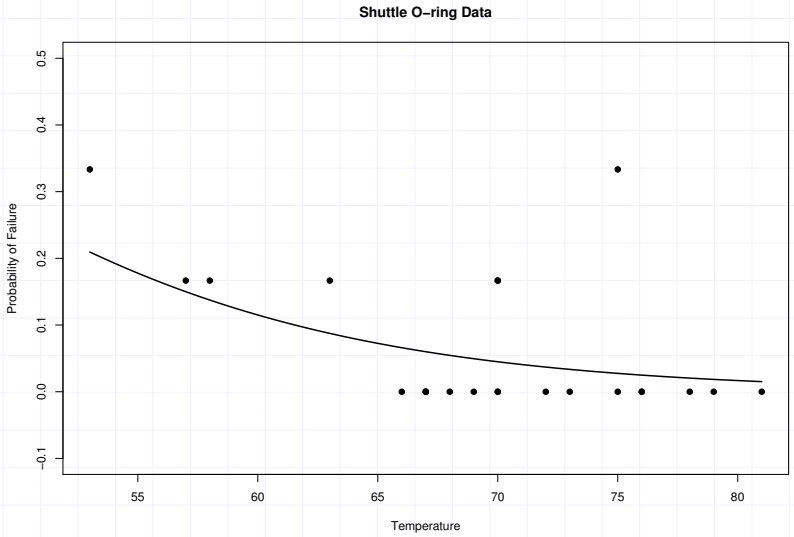
```
oring=read.csv("~/joe/lectures/DT9209\ MSc\ Statistics/Data/
    oring.csv",header=T)
attach(oring)
m=matrix(c(nFailures,6-nFailures),ncol=2,byrow=F)
fit1=glm(m~Temperature,family=binomial())
```

Note that glm() will fit other exponential family regression models, even normal data, hence the 'family' option needs to be specified.

The results are given as table of means and SEs as well as some other statistics.

```
summary(fit1)

Call:
glm(formula = m ~ Temperature, family = binomial())

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-0.95227  -0.78299  -0.54117  -0.04379   2.65152

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.08498    3.05247   1.666   0.0957 .
Temperature -0.11560    0.04702  -2.458   0.0140 *
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24.230  on 22  degrees of freedom
Residual deviance: 18.086  on 21  degrees of freedom
AIC: 35.647
```


The fit of this model may be plotted.



Properties of the MLE

What are the properties of the Maximum Likelihood Estimates (MLEs) from a GLM?

To answer this we need a result from ML theory.

Definition 1 (score function): The score function $U(\beta)$ is the first derivative of the log likelihood with regard to the parameter vector β, i.e.

$$\log L(\beta; Y) = \sum_{i=1}^{n} \log f(Y; \beta)$$

$$U(\beta) = \frac{\log L(\beta; Y)}{\partial \beta}$$

At the maximum likelihood estimate $\hat{\beta}$, $U(\hat{\beta}) = 0$ since it is at the maximum of the function.

Definition 2 (observed Fisher information): The first derivative of the score function WRT β evaluated at the maximum, is a measure of the curvature of the likelihood around the maximum. The greater the curvature, the more information in the data as to the location of the maximum. The curvature captures the precision (information) of the estimated parameters in a sense.

The first derivative of the score function is the second derivative of the likelihood,

$$U'(\beta) = \frac{\partial U(\beta)}{\partial \beta} = H = \frac{\log L^2(\beta; Y)}{\partial \beta^2}$$

This H is the Hessian from above (will be scalar when β is scalar) or in general $p \times p$ where there are p regression parameters.

This will be negative when evaluated at the maximum $\hat{\beta}$, so take the negative and call the result the observed Fisher information, $I(\beta)$:

$$I(\beta) = -H = -\frac{\log L^2(\beta; Y)}{\partial \beta^2}$$

With some technical results it can be shown that asymptotically the distribution of the MLE is given by,

$$\hat{\beta} \sim N(\beta, I^{-1}(\hat{\beta}))$$

Therefore the standard error of the MLE of a parameter is given by,

$$se(\hat{\beta}_j) = I_{jj}^{-\frac{1}{2}}(\hat{\beta}) \quad (4)$$

i.e. the observed information matrix $(I^{-1}(\hat{\beta}))$ is an approximate variance-covariance for the parameter estimates. So, the appropriate diagonal of the observed information matrix is the variance of the individual parameter.

For Normal data this result is exact.

NB. These results hold asymptotically. That means you need a lot of data to be confident that the results do in fact hold.
In practice, for the exponential family, once there is a reasonable amount of data then the asymptotic results hold well. If in doubt, then a simple check is to plot the likelihood and the asymptotic normal approximation (which represents a quadratic approximation to the likelihood around the MLE) and see if they agree.
E.G. see, Pawitan, 2001, 'In all likelihood: statistical modelling and inference using likelihood', Oxford University Press, Oxford

From all this we can derive a practical hypothesis testing procedure and method for calculating CIs - the Wald test and interval.

To test $H_0 : \beta = \beta^0$ versus $H_a : \beta \neq \beta^0$ use,

$$z = \frac{\hat{\beta} - \beta^0}{se(\hat{\beta})} \sim N(0, 1)$$

or

$$\chi^2 = \frac{(\hat{\beta} - \beta^0)^2}{se^2(\hat{\beta})} \sim \chi^2_1 \quad (5)$$

A Wald based CI may be derived as;

Wald based CI: $$\hat{\beta} \pm z_{1-\alpha/2} se(\hat{\beta}) \quad (6)$$

Example: Test and CI for the Temperature parameter - O-ring data.
From the N-R algorithm we get at convergence;

$$\begin{pmatrix} 5.0850 \\ -0.1156 \end{pmatrix} = \begin{pmatrix} 5.0850 \\ -0.1156 \end{pmatrix} - \begin{pmatrix} -7.93 & -511.57 \\ -511.57 & -33434.56 \end{pmatrix}^{-1} \begin{pmatrix} 0.00 \\ 0.00 \end{pmatrix}$$

So we have,

$$\hat{\beta} = \begin{pmatrix} 5.0850 \\ -0.1156 \end{pmatrix} \qquad H = \begin{pmatrix} -7.93 & -511.57 \\ -511.57 & -33434.56 \end{pmatrix}$$

$$\Rightarrow I(\hat{\beta}) = \begin{pmatrix} 7.93 & 511.57 \\ 511.57 & 33434.56 \end{pmatrix}$$

$$I^{-1}(\hat{\beta}) = \begin{pmatrix} 9.3177 & -0.1426 \\ -0.1426 & 0.0022 \end{pmatrix}$$

1. Test the null hypothesis that the slope is zero.
2. Calculate a 95% CI for the slope,
3. Calculate a 95% CI for the odds ratio for a 1 degree rise in temperature.

Model Building and Likelihood Ratio Tests

The same model building issues for GLMs arise as we looked at for linear models. In particular the issue of type I and type III tests arise again - but this time we measure not SS but **deviance**.

Definition (Deviance): Deviance is defined as,

$$-2 \log \frac{\hat{L}_c}{\hat{L}_{max}} = -2 \log \hat{L}_c + 2 \log \hat{L}_{max}$$

where $\hat{L}_c$ is the value of the likelihood at the maximum for the current model under consideration and $\hat{L}_{max}$ is the maximum possible likelihood.

The maximum possible log likelihood is given by the saturated model - i.e. there is a parameter for every observation in the data set - a perfect fit but with no simplification.

This means for a binomial case that the observed proportion at a given value of the covariates is matched perfectly by the fitted value.

The saturated model for binomial data has parameter $\hat{\pi}_i = y_i/n_i$ for each observation i. Then the value of $\log \hat{L}_{max}$ is:

$$\log \hat{L}_{max} = \sum_{i=1}^{n} \left\{ \log \binom{n_i}{y_i} + y_i \log(y_i/n_i) + (n_i - y_i) \log[1 - y_i/n_i] \right\}$$

and the value of $\hat{L}_c$ is the exponential of,

$$\log \hat{L}_c = \sum_{i=1}^{n} \left\{ \log \binom{n_i}{y_i} + y_i \log(\hat{p}_i) + (n_i - y_i) \log[1 - \hat{p}_i] \right\}$$

where $\hat{p}_i$ is the fitted probability for the value of covariates x_i' for that i^{th} observation.

It can be shown that the distribution of the deviance statistic is approximately χ^2 with $n - p$ degreees of freedom where n is the number of binomial obervations in the data and p is the number of parameters in the current model.

This approximation may be used to assess the goodness of fit of the model to the data. However, the total number of binary observation must be 'large' for the approximation to be sufficiently good for this purpose.

A better use for the deviance is to compare **nested** models. Two models are nested where the all covariates in one model are also to be found in the other model, e.g.

$$\begin{array}{lrcl} \text{Model 1:} & \text{logit}(p_i) & = & \beta_0 \\ \text{Model 2:} & \text{logit}(p_i) & = & \beta_0 + \beta_1 x_1 \\ \text{Model 3:} & \text{logit}(p_i) & = & \beta_0 + \beta_2 x_2 \\ \text{Model 4:} & \text{logit}(p_i) & = & \beta_0 + \beta_1 x_1 + \beta_2 x_2 \end{array}$$

Models 1-3 are nested in model 4.

Model 1 in nested in model 2, 3 and 4.

Model 2 is not nested in model 3 as the x_1 covariate in model 2 is not also in model 3.

Model 3 is not nested in model 2 as the x_2 covariate in model 3 is not also in model 2.

For 2 models a and b, where a is nested in b it can be shown that under the null hypothesis, $D_a - D_b$ is approximately distributed as a χ^2 random variable with degrees of freedom $p_b - p_a$, where $D_a,\ D_b$ are the deviance for model 1 and 2 respectively, $p_a,\ p_b$ are the number of parameters in model a and b respectively. This can be used as the basis for model selection.

Nb: $D_a - D_b = -2\log$ lik. (Model a) $+2\log$ lik. (model b)

Neuralgia data

Treatment	Sex	Age	Duration	Pain	Treatment	Sex	Age	Duration	Pain
P	F	68	1	0	P	F	79	20	1
B	M	74	16	0	A	M	70	12	0
P	F	67	30	0	A	F	69	12	0
P	M	66	26	1	B	F	65	14	0
B	F	67	28	0	B	M	70	1	0
B	F	77	16	0	B	M	67	23	0
A	F	71	12	0	A	M	76	25	1
B	F	72	50	0	P	M	78	12	1
B	F	76	9	1	B	M	77	1	1
A	M	71	17	1	B	F	69	24	0
A	F	63	27	0	P	M	66	4	1
A	F	69	18	1	P	F	65	29	0
B	F	66	12	0	P	M	60	26	1
A	M	62	42	0	A	M	78	15	1
P	F	64	1	1	B	M	75	21	1
A	F	64	17	0	A	F	67	11	0
P	M	74	4	0	P	F	72	27	0
A	F	72	25	0	P	F	70	13	1
P	M	70	1	1	A	M	75	6	1
B	M	66	19	0	B	F	65	7	0
B	M	59	29	0	P	F	68	27	1
A	F	64	30	0	P	M	68	11	1
A	M	70	28	0	P	M	67	17	1
A	M	69	1	0	B	M	70	22	0
B	F	78	1	0	A	M	65	15	0
P	M	83	1	1	P	F	67	1	1
B	F	69	42	0	A	M	67	10	0
B	M	75	30	1	P	F	72	11	1
P	M	77	29	1	A	F	74	1	0
A	F	69	3	0	B	M	80	21	1

A study was conducted of the analgesic effects of treatments on elderly patients with neuralgia. Compare two treatments with a placebo, but controlling for age, gender and disease duration.

To fit the full model is R we use the code;

```
neuralgia=read.table("~/joe/lectures/DT9209\ MSc\ Statistics
    /Data/neuralgia.txt",header=T,sep='')
fit2=glm(Pain~Age+Duration+factor(Sex)+factor(Treatment),
    family=binomial(),data=neuralgia)
drop1(fit2,test='LRT')
```

The data is not grouped so the binomial divisor here is 1 - and so can be omitted.

The output from R is;

```
> summary(fit2)
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -20.588282   7.102883  -2.899  0.00375 **
Age                  0.262093   0.097012   2.702  0.00690 **
Duration            -0.005859   0.032992  -0.178  0.85905
factor(Sex)M         1.832202   0.796206   2.301  0.02138 *
factor(Treatment)B  -0.526853   0.937025  -0.562  0.57394
factor(Treatment)P   3.181690   1.016021   3.132  0.00174 **
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 81.503  on 59  degrees of freedom
Residual deviance: 48.736  on 54  degrees of freedom
AIC: 60.736

Number of Fisher Scoring iterations: 5
```


Notice that the set-to-zero constraint here is the same as for lm().

The interpretation of the parameter for a classification variables is the log odds ratio for being at that level of the classification variable versus the level set to zero - but with all other variables held the same.

What is the correct interpretation of the parameter for treatment B?

To determine if all the variables included in this model LRTs can be performed.

Model No.	Model	p	log lik	-2 log lik	AIC
1	age	2	-36.53	73.06	77.06
2	sex	2	-37.92	75.84	79.84
3	duration	2	-39.94	79.88	83.88
4	treat	3	-33.74	67.48	73.48
5	age sex	3	-34.45	68.90	74.90
6	age duration	3	-36.24	72.48	78.48
7	age treat	4	-27.52	55.04	63.04
8	sex duration	3	-37.17	74.34	80.34
9	sex treat	4	-29.94	59.88	67.88
10	duration treat	4	-33.34	66.68	74.68
11	age sex duration	4	-34.21	68.42	76.42
12	age sex treat	5	-24.38	48.76	58.76
13	age duration treat	5	-27.52	55.04	65.04
14	sex duration treat	5	-29.61	59.22	69.22
15	age sex duration treat	6	-24.37	48.74	60.74

There are a number of model building strategies - forward, backward and stepwise as before.

One way to go is as follows:

Test for treat; compare model 15 with 11. Conclude?

Test for duration: compare 15 with 12. Conclude?

Test ?: compare 12 with 5. Conclude?

test ?: compare 12 with 7. Conclude?

test ?: compare 12 with 9. Conclude?

Information Criteria and Model building

There are a number of Information Criteria that are routinely used in model building. The most important is the Akaike Information Criterion (AIC).

AIC is based on fairly complex theory from information entropy - so we will omit any deep discussion. It takes a remarkably simple form however:

AIC $= -2$ log likelihood (evaluated at MLE) $+2p$

In this version the model with smaller AIC is preferred, therefore models with more parameters are penalised.

This is helpful, as the log likelihood is a non-decreasing function of the number of parameters.

The main advantage of the AIC is that is can be used to compare non-nested models - with the model with the smaller AIC preferred.

Drawback to AIC are:

(1) we don't know its distribution

(2) it can be misleading in small samples.

In the case of drawback (2) we can use the corrected AIC or AICc:

AICc $= -2$ log likelihood (evaluated at MLE) $+2p\frac{n}{n-p-1}$

The Exponential Family

Linear regression is a special case of GLM. It is where the distribution of the data is assumed to be normal.

In general, GLM is concerned with regression type models for a class of probability distributions called the exponential family.

The exponential family of distributions are those probability distributions that can be parameterised in the following way;

$$f(y;\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{7}$$

For some function $a(.)$,$b(.)$,$c(.)$ and for parameters θ and ϕ. If ϕ is known then this is exponential family model with **canonical parameter** θ (also called the natural parameter). If ϕ is unknown then this may be a two parameter exponential family.

Example: Normal Distribution

$$\begin{aligned} f(y;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}}\, exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \\ &= exp\left\{\frac{y\mu-\mu^2/2}{\sigma^2} - \frac{1}{2}\left[y^2/\sigma^2 + \log(2\pi\sigma^2)\right]\right\} \end{aligned}$$

where $\theta = \mu$, $\phi = \sigma^2$, $a(\phi) = \phi$, $b(\theta) = \theta^2/2$ and
$c(y,\phi) = -\frac{1}{2}\{y^2/\sigma^2 + \log(2\pi\sigma^2)\}$

Example: Poisson Distribution

$$f(y;\lambda) = \frac{\lambda^y exp(-\lambda)}{y!} = exp\left\{y\log\lambda - \lambda - \log(y!)\right\}$$

where $\theta = \log\lambda$, $\phi = 1$, $a(\phi) = 1$, $b(\theta) = exp(\theta)$ and
$c(y,\phi) = -log(y!)\phi$

Example: Binomial Distribution

$$\begin{aligned} f(y;\pi) &= \binom{n}{y}\pi^y(1-\pi)^{n-y} \\ &= exp\left\{\log\binom{n}{y} + y\log\pi + n\log(1-\pi) - y\log(1-\pi)\right\} \\ &= exp\left\{y\log\left(\frac{\pi}{1-\pi}\right) + \log\binom{n}{y} + n\log(1-\pi)\right\} \end{aligned}$$

where $\theta = \log(\frac{\pi}{1-\pi})$, $\phi = 1$, $a(\phi) = 1$, $b(\theta) = n\log(1+e^\theta)$ and
$c(y,\phi) = \log\binom{n}{y}\phi$

Some other members of the exponential family of distributions are:

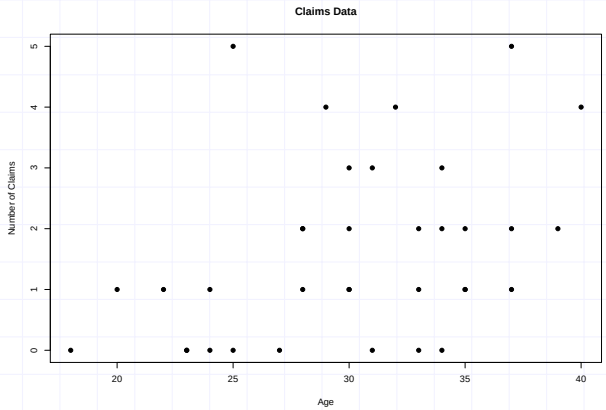
- Exponential distribution
- Negative binomial (with known number of failures
- χ^2
- Gamma distribution
- Multinomial (with known number of trials).
- many others....

Poisson regression

A sample of data on the number of health insurance claims by customers of different ages in give in the table below.

Customer ID	Age	Number of Claims	Customer ID	Age	Number of Claims
1	18	0	19	31	0
2	20	1	20	31	3
3	22	1	21	32	4
4	23	0	22	33	2
5	23	0	23	33	0
6	24	0	24	33	1
7	24	1	25	34	2
8	25	0	26	34	3
9	25	5	27	34	0
10	27	0	28	35	1
11	28	1	29	35	2
12	28	2	30	35	1
13	28	2	31	37	2
14	29	4	32	37	5
15	30	2	33	37	1
16	30	1	34	39	2
17	30	3	35	40	4
18	30	1			

* Data from Pawitan, 2001



NB. (1) The responses are integers, (2) the variance seems to increase with the mean and, (3) the number of claims has no natural limit.

An obvious choice for distribution with these characteristics in the Poisson.

$$p(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}, \qquad \lambda > 0, \qquad y \in \{0, 1, 2, 3, \ldots\}$$

The $E[Y] = \lambda$ and $Var[Y] = \lambda$

Poisson regression Model

We want to relate age to the mean response. The mean response must be positive so a natural choice is the following:

$$E[Y_i] = \lambda_i = e^{x_i'\beta} \qquad \Rightarrow \log \lambda_i = x_i'\beta$$

Which leads to the following likelihood and log likelihood (assuming independence between observations):

$$\begin{aligned} L(\beta; Y, X) &= \prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \\ \ell(\beta; Y, X) &= \sum_{i=1}^{n} -e^{x_i'\beta} + y_i x_i'\beta \end{aligned} \tag{8}$$

Give that for the health claims data, $x_i'\beta = \beta_0 + \beta_1(age_i)$, find the score vector and hessian matrix.

Hence, iterate to solution....[exercise]

The correct solution using R is given by:

```
1 claims=read.table("~/joe/lectures/DT9209\ MSc\ Statistics/
      Data/claims.txt",header=T,sep='')
2 fit3=glm(claims~age,family=poisson(),data=claims)
3 summary(fit3);logLik(fit3)
```

Interpret both $\hat{\beta}_1$ and $\hat{\beta}_0$?

Hypothesis testing can proceed as for the logistic regression case, i.e. using Wald or (better) LR based tests.

Model building can be based on hypothesis testing using stage-wise algorithms, and/or AIC/AICc etc.

Fitted values from Logistics & Poisson Regression

A general linear hypothesis of the parameters can be applied to all GLMs, including logistic and Poisson regression models.

These take the form of:

$$L\beta = 0$$

We can use Wald based tests/intervals in such cases, i.e. based on:

$$L\hat{\beta} \sim N\left(L\beta, LI^{-1}(\hat{\beta})L'\right)$$

This is conceptually very close to the GLH we used in linear regression. LR based hypotheses can also be used and are theoretically better for finite samples.

We can use these to calculate CI's for fitted values on the linear predictor scale- simply replace L with a suitably specified x_i':

$$CI_{(1-\alpha)\%} = x_i'\hat{\beta} \pm z_{(1-\alpha)\%}\sqrt{x_i' I^{-1}(\hat{\beta}) x_i}$$

and finally either exponentiate both limits (Poisson) or using the logit inverse on the limits (Logistics regression).

Poisson regression with Offset

Epilepsy data

Subject	treatment	Weeks	Attacks
1	active	12	3
2	active	5	2
3	active	7	4
4	active	14	3
5	active	10	5
6	active	10	2
7	active	12	1
8	active	8	3
9	active	11	3
10	active	8	3
11	placebo	11	4
12	placebo	11	7
13	placebo	8	6
14	placebo	16	8
15	placebo	11	11
16	placebo	7	8
17	placebo	15	7
18	placebo	9	7
19	placebo	7	4
20	placebo	4	2
21	placebo	6	6
22	placebo	4	1

- The variable follow-up time is a crucial feature of these data
- The Poisson mean is a rate in unit time - however the unit is defined.
- Therefore we need to use an **offset** term in the Poisson model to account for this.
- Such an offset term is used whenever the number of events observed is recorded from observations with different exposure to the event.

Model including Offset

$$\lambda_i = t_i e^{x_i'\beta} = e^{\log t_i + x_i'\beta}$$

where t_i is the exposure for observation i.

This leads to the following log likelihood:

$$\ell(\beta; Y, X, T) = \sum_{i=1}^{n} -e^{\log t_i + x_i'\beta} + y_i x_i'\beta \qquad (9)$$

This log likelihood is maximised as before, WRT to the β's - NB. not with respect to exposure.

In essence, the exposure term acts as a weight for each observation in the maximisation.

We implement in the software by specifying the exposure variable as a predictor with a known coefficient of one.

```
1 epilepsy=read.table("~/joe/lectures/DT9209\ MSc\ Statistics/
      Data/epilepsy.txt",header=T,sep=',')
2 fit4=glm(Attacks~factor(treatment)+offset(log(Time.)),family
      =poisson(),data=epilepsy)
3 summary(fit4);logLik(fit4)
4 #### see what happens without offset
5 fit4a=glm(Attacks~factor(treatment),family=poisson(),data=
      epilepsy)
6 summary(fit4a);logLik(fit4a)
```