GENERALIZED LINEAR MODELS

Introduction (1)

1. Definition of GLM and the use of maximum likelihood (ML) based inference in the context of GLM
2. The main class of GLM and their relevance in medical and epidemiological questions
3. The interpretation of parameter’s from GLM’s
4. The use of Stata to model data with GLM
5. Comparison and assessment of fit of GLM’s
References

Modeling process

1. The model

A statistical model is a representation of the population under study. The model also usually reflects the way in which data have been sampled from the population.

2. Parameters

Parameters in the model correspond to features of the population. In practice the true values of the parameters are unknown and so the sample is to be used to draw inferences about them.
Modeling process (continue)

3. Parameter estimators

There are sample statistics that are used to provide estimates of the unknown parameters. Recall that estimator refers to the general form of the statistic (a random variable), while estimate refers to the actual numerical realization from a given sample.

4. Hypotheses

There are statements about the unknown parameters, e.g. a parameter takes a particular values $\pi=1/2$ or two parameters are equal $\pi_1 = \pi_2$
MODEL
Statistical models contain both: **Systematic** effects and **Random** effects

regression models:

\[ Y_i = \beta_0 + \sum \beta_j x_{ij} + e_i \]

Discussion:
1) What sort of variable \( Y \) is (continuous, discrete, qualitative, …)?
2) What is the distribution of \( Y \), what is the range of possible values?
3) Does the sample of observed \( Y \)’s have to fit the distribution exactly? What else is important about the \( Y \)’s distribution?
4) What sort of variables are the \( x \)’s? (continuous, discrete, qualitative)
5) How do we assess how well this model fits a set of data?
6) How do we assess how well individual cases conform to the fitted model?
7) How do we assess the effect of individual cases on the fitted model?
BINOMIAL DISTRIBUTION

Y_i’s binomial.
The response variable will be Y_i = 0 or 1 (ungrouped data)
Case control study:

Y_i=0 for control
Y_i=1 for case.

Grouped data: A group of n_i cases all have the same values
of x_ij’s. In that case
Y_i=# cases; Y_i=0,1, ..., n_i
Proportion =Y_i/n_i.

Can we use ordinary regression with binomial Y_i?

Ungrouped data: NO
Grouped data: In some cases YES

Potential problems:
Model is actual

\[
Y_i / n_i = \beta_0 + \sum \beta_j x_{ij} + e_i
\]

1. Y_i/n_i bounded 0…1
2. Variance of Y_i/n_i not constant (but may be approximately
   proportional to 1/n_i).

We can use binomial distribution
BASIC GLM’s ASSUMPTIONS

1) *Independent observations*

More generally, the observations are independent in blocks of fixed known sizes.

*Consequence:* Data exhibiting the autocorrelations of time series are *excluded.*

2) *There is a single error term in the model*

This constraint excludes, for example models for the analysis of experiments having more than one error term (split-plot design i.e. between and within-plot variance).
The concept of “Likelihood”

If \( Y_1, \ldots, Y_n \) are independent random variables each with probability density function (pdf) \( f_i(y_i; \theta) \), where \( \theta \) is a (possibly vector-valued) parameter, then, by virtue of independence, the joint pdf of the vector \( Y \) is

\[
L(\theta; y) = \prod_{i=1}^{n} f_i(y_i; \theta)
\]

This function, considered a function of the parameter \( \theta \), is called the \textit{likelihood}.

Usually we work with the logarithm of the likelihood function and under the assumption of independence of the observations we have

\[
l(\theta; y) = \log L(\theta; y) = \sum_{i=1}^{n} \log f_i(y_i; \theta)
\]
**Maximum likelihood estimation**

A way to estimate $\theta$, is by finding a value $\hat{\theta}$ such that

$$L(\hat{\theta}; y) \geq L(\theta; y), \forall \theta \in \Theta$$

where $\Theta$ is the space of $\theta$. This value $\hat{\theta}$ is called the **maximum-likelihood estimate** (MLE) of $\theta$.

We can more easily calculate $\theta$ by maximizing the log-likelihood. The MLE of $\theta$ also maximizes this function. That is,

$$l(\hat{\theta}; y) \geq l(\theta; y), \forall \theta \in \Theta$$

Working with the log-likelihood is preferred because it is easier to maximize sums of functions versus products.
MLE’s (continue)

1. For regular problems, $\hat{\theta}$ can be obtained by equating the first derivative of the likelihood function (or equivalently of the log-likelihood) to zero. Provided the second derivative at this point is negative, the resulting value is the MLE.

2. The likelihood is a function of the parameters and we are interested in its behavior (or shape) with respect to them. Constant multiplicative (additive) terms not involving the parameters can be dropped.

3. MLE’s have a number of important properties that make them desirable.
   a) MLE’s are asymptotically unbiased i.e. the expectation of $\hat{\theta}$, $E(\hat{\theta})$, becomes equal to $\theta$ as $n \to \infty$.
   b) A MLE has a sampling distribution that is asymptotically normal with variance the inverse of minus the information ($\{d^2 l / d\theta^2\}^{-1}$).
   c) MLE’s are invariant under transformation i.e. if $\hat{\theta}$ is the MLE of $\theta$ then any function of $\hat{\theta}$ will be the MLE of the same function of $\theta$. 
The Binomial distribution

Consider for example the Binomial distribution, of counts \( y_i, i=1,\ldots,n \) and each \( y_i=0,1,2,\ldots,n_i \).

\[
f_i(y_i; \pi_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i-y_i}
\]

The likelihood function is, from above,

\[
L(\theta; y) = \prod_{i=1}^{n} f_i(y_i; \theta)
\]

that is,

\[
L(\theta; y) = \prod_{i=1}^{n} f_i(y_i; \pi_i) = \prod_{i=1}^{n} \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i-y_i}
\]

Maximizing this likelihood involves taking derivatives with respect to \( \theta=(\pi_1,\ldots,\pi_n) \).
The Binomial distribution (continued)

Consider the log likelihood

\[ l(\theta; y) = \sum_{i=1}^{n} \log f(y_i; \pi_{i}) \]

\[ = \sum_{i=1}^{n} \log \left( \frac{n_i}{y_i} \pi_{i}^{y_i} (1 - \pi_{i})^{n_i - y_i} \right) \]

\[ = \sum_{i=1}^{n} \left( y_i \log \pi_{i} + (n_i - y_i) \log (1 - \pi_{i}) + \log \left( \frac{n_i}{y_i} \right) \right) \]

\[ = \sum_{i=1}^{n} y_i \log \pi_{i} + \sum_{i=1}^{n} (n_i - y_i) \log (1 - \pi_{i}) + \sum_{i=1}^{n} \log \left( \frac{n_i}{y_i} \right) \]

\[ = \log \pi_{i} \sum_{i=1}^{n} y_i - \log (1 - \pi_{i}) \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} n_i \log (1 - \pi_{i}) + C \]
Maximum-likelihood estimation

To maximize the above expression, we must take $n$ derivatives with respect to $\pi_i$, set them all equal to zero and solve a system of $n$ equations with $n$ unknowns. Let’s consider the much simpler case, where $\pi_1 = \pi_2 = \cdots = \pi_n = \pi$.

Then taking a derivative with respect to $\pi$ becomes,

$$
\frac{dll(\pi; y)}{d\pi} = \frac{1}{\pi} \sum_{i=1}^{n} y_i + \frac{1}{1-\pi} \sum_{i=1}^{n} y_i - \frac{1}{1-\pi} \sum_{i=1}^{n} n_i = 0
$$

which becomes

$$
\frac{\sum_{i=1}^{n} y_i - \pi N}{\pi(1-\pi)} = 0 \Leftrightarrow \sum_{i=1}^{n} y_i = \pi N
$$

since $\pi \in (0,1)$ and finally,

$$
\hat{\pi} = \frac{1}{N} \sum_{i=1}^{n} y_i = \bar{y}
$$
**Example:** Consider the situation \( n_1 = n_2 = n_3 = n_4 = n_5 = n \) and \( y_1 = 2, y_2 = 1, y_3 = 1, y_4 = 3, y_5 = 3 \). A plot of the log-likelihood is as follows:

The log likelihood is maximized at \( \hat{\pi} = \bar{y} = 0.2 \).
**Linear regression**

Suppose that two measurements \((y_i, x_i)\) are made on each of \(n\) individuals. A simple linear regression model of \(y\) on \(x\) can be written:

\[
y_i = \alpha + \beta x_i + e_i
\]

where \(e_i\) are independent with \(e_i \sim N(0, \sigma^2)\).

Alternatively we can say that the random variable \(Y_i\) has a conditional distribution

\[
Y_i \mid x_i \sim N(\alpha + \beta x_i, \sigma^2)
\]

This shows more clearly what is actually happening in a regression model. In general, in any regression model, the **conditional expectation** of a random variable (say \(Y\)) **given** the values of one or more other variables \((x=\{x_1, x_2, \ldots, x_p\})\) is expressed as some function of these fixed variables (covariates, independent variables) and some parameters \((\beta = \{\beta_1, \beta_2, \ldots, \beta_p\}^T)\)

\[
E(Y \mid X) = g(X, \beta)
\]
Linear regression (continue)

The log-likelihood can be written:

\[
l(\alpha, \beta, \sigma^2 \mid Y, X) = -\frac{n}{2} \ln(2\pi \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2
\]

The MLE’s of \( \alpha \) and \( \beta \) can be obtained from the score equations:

\[
U(\alpha) = l'(\alpha) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)
\]

\[
U(\beta) = l'(\beta) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) x_i
\]

\[
U(\alpha) = 0 \Rightarrow \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}
\]

\[
U(\beta) = 0 \Rightarrow \hat{\beta} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2}
\]

Note that we do not need to know \( \sigma^2 \) to solve these. These estimates are equivalent to the least squares estimators derived by minimizing the residual sum of squares. Hence, in this case, maximum likelihood is equivalent to least squares.
The exponential family of distributions

In GLMs the observations are assumed to arise from the exponential family of distributions

\[ f(y; \theta, \phi) = \exp \left\{ y \theta - b(\theta) / a(\phi) + c(y, \phi) \right\} \]

where \( a(\phi), b(\theta) \) and \( c(y, \phi) \) are known functions. The parameter \( \theta \) is known as the canonical parameter. In general, it can be shown that

\[ E(Y) = b'(\theta) \quad \text{and} \quad \text{var}(Y) = b''(\theta) a(\phi) \]

The variance is thus a product of two terms, \( b''(\theta) \) which depends on the mean (through \( \theta \)) which is called the variance function \( V(\mu) \), and the other on \( a(\phi) \), a function of the form \( a(\phi) = \phi / \omega = \sigma^2 / \omega \)

\( \phi \) is called the dispersion or scale parameter, is constant over observations and \( \omega \) known prior weights that vary from observation to observation.
Examples: The Normal distribution

The density of a $N(\mu, \sigma^2)$ random variable $Y$ can be written:

$$f(y) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} (y - \mu)^2 \right\}$$

and the logarithm of it is

$$\ln\{f(y)\} = -\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2} \ln(2\pi\sigma^2) =$$

$$= (y\mu - \frac{\mu^2}{2})/\sigma^2 - \frac{1}{2} \left\{ \frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2) \right\}$$

so,

$$\theta = \mu, \quad b(\theta) = \frac{\theta^2}{2}, \quad a(\phi) = \alpha = \sigma^2, \quad and c(y, \phi) = \left\{ \frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2) \right\}$$

$$E(y) = b'(\theta) = \theta = \mu$$

The mean of the normal distribution it is equal to the canonical parameter $\theta$

$$\text{var}(Y) = b''(\theta)a(\phi) = \phi = \sigma^2$$

of the form $\alpha(\phi) = \phi / \omega$ with prior weights equal to 1.
Example: The Poisson distribution

The density of a $P(\lambda)$ random variable $Y$ can be written:

$$f(y) = \Pr(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!} \quad y = 0, 1, 2, ...$$

and the logarithm of it is

$$\ln\{f(y)\} = y \ln(\lambda) - \lambda - \ln(y!)$$

so,

$$\theta = \ln(\lambda) \quad b(\theta) = \lambda = e^\theta \quad a(\phi) = 1 \quad and \quad c(y, \phi) = \ln(y!)$$

$$E(y) = b'(\theta) = e^\theta = \lambda$$

The mean of the Poisson distribution

$$\text{var}(Y) = b''(\theta)a(\phi) = e^\theta * 1 = \lambda$$
**Example: The Binomial distribution**

The density of a $B(n, \pi)$ random variable $Y$ can be written:

$$f(y) = \Pr(Y = y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y} \quad y = 0, 1, 2, \ldots, n$$

and the logarithm of it is

$$\ln\{f(y)\} = y \ln(\pi) + (n - y) \ln(1 - \pi) + \ln\left\{\binom{n}{y}\right\} =$$

$$= y \ln\left(\frac{\pi}{1 - \pi}\right) + n \ln(1 - \pi) + \ln\left\{\binom{n}{y}\right\}$$

so,

$$\theta = \ln\left(\frac{\pi}{1 - \pi}\right) \quad , \quad b(\theta) = -n \ln(1 - \pi) = n \ln(1 + e^\theta)$$

$$a(\phi) = 1 \quad \text{and} \quad c(y, \phi) = \ln\left\{\binom{n}{y}\right\}$$

$$E(y) = b'(\theta) = n \frac{e^\theta}{1 + e^\theta} = n \pi$$

The mean of the Binomial distribution

$$\text{var}(Y) = b''(\theta)a(\phi) = n \frac{e^\theta}{(1 + e^\theta)^2} = n \pi (1 - \pi)$$
Normal linear regression

To introduce GLMs we turn to the familiar general linear model. There, the observations $y_1, y_2, \ldots, y_n$ are distributed according to normal distributions $N(\mu_i, \sigma^2)$.

More importantly, we have that the expected value $\mu_i = E(y_i)$ is related to a $p$-dimensional vector $x'_i = (x_{i1}, \ldots, x_{ip})$ as follows:

$$\mu_i = x'_i \beta = \sum_{j=1}^{p} x_{ij} \beta_j$$

with $\beta$ being an unknown $p$-dimensional parameter vector.

The components of $Y$ are independent Normal variates with variance $\sigma^2$ and

$$E(Y) = \mu \quad \text{where} \quad \mu = X\beta \quad (1)$$
The generalization

1) **The random part:** the components of $Y$ have independent *Normal distributions* with $E(Y) = \mu$ and constant variance $\sigma^2$

2) **The systematic component:** covariates $x_1, x_2, \ldots, x_p$ produce a linear predictor $\eta$ given by

$$\eta = \sum_{j=1}^{p} x_{ij} \beta_j$$

3) The **link** between the random and the systematic components:

$$\mu = \eta$$

This generalization introduces a new symbol $\eta$ for the *linear predictor* and the third component that specifies that $\mu$ and $\eta$ are in fact identical. If we write

$$n_i = g(\mu_i)$$

then $g(.)$ is called the **link function**
Generalization (continue)

The association of the mean $\mu_i = E(y_i)$ with $x_i$ is modeled by introducing a link function $g$

$$\eta_i = g(\mu_i)$$

where $\eta_i$ is a linear combination

$$\eta_i = x_i^t \beta = \sum_{j=1}^{p} x_{ij} \beta_j$$

**Canonical link:** The canonical parameter is equal to the linear predictor

$$\theta = \eta = \sum_{j=1}^{p} x_{ij} \beta_j$$

The canonical links have attractive statistical properties because in this case there exists a sufficient statistic

$$\sum_{i} x_{ij} y_{ij}, j = 1, \ldots, p$$
Canonical links

Normal \[ \eta = \mu \quad \text{Identity} \]

Poisson \[ \eta = \log(\mu) = \log(\lambda) \quad \text{Log of rate} \]

Binomial \[ \eta = \log(\pi/(1 - \pi)) \quad \text{Logit} \]
Comments on GLM’s

1. The linear predictor is linear in the parameters, not necessarily in the explanatory variables.

   *Polynomial regression:*

   $\eta_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p$

2. The ‘log likelihood of the model’ is usually shorthand for ‘the log likelihood function of the model for the given data evaluated at the MLE’s of the parameters’ i.e. the maximum of the log likelihood function

3. With minor exceptions the underlying theory (parameter estimation, inference, model assessment and comparison) proceeds in the same way for all GLM’s; it is just the distribution and the link that differ.

4. The link function is chosen to provide a suitable scale for the effects of explanatory variables to operate in a linear manner. Typically the range of $\mu$ will be transformed to the whole real line ($-\infty$ to $\infty$).
**Likelihood-ratio tests**

Two models are called *nested* if one is “contained” in some sense in the other.

The likelihood-ratio test compares the maximized likelihood of the two nested models, i.e.,

$$
\lambda = \log \frac{L(\hat{\theta}_1; y)}{L(\hat{\theta}_2; y)} = l(\hat{\theta}_1; y) - l(\hat{\theta}_2; y)
$$

where $\hat{\theta}_1$ and $\hat{\theta}_2$ are the MLEs in the smaller and larger models respectively.

$\lambda$ takes values from 0 to 1 with lower values favoring the larger model.

In large samples, $-2 \log \lambda \sim \chi^2_v$, where $v$ is the difference in dimension in the two models.
Example: Binomial distribution

Consider the example: Binomial, \( n_1 = n_2 = n_3 = n_4 = n_5 = 10; \ y_1 = 2, \ y_2 = 1, \ y_3 = 1, \ y_4 = 3, \ y_5 = 3; \hat{\pi} = 0.2 \)

Hypothesis \( H_0 : \pi = 0.3 \)

We compare the log-likelihood of the binomial model under the null hypothesis

\[
l(0.3; y) = -26.307
\]

while its maximum, evaluated at the MLE \( \hat{\pi} = 0.20 \)

\[
l(0.2; y) = -25.020.
\]

Likelihood ratio under the two models

\[
-2\log \lambda = -2(-26.307 + 25.020) = 2.573
\]

which is asymptotically distributed as a chi-square distribution with one degree of freedom.

The p-value = 0.109 which is non-significant
Deviance

**Saturated** model: the one with the maximum possible number of parameters, with no redundancies i.e, the same number of parameters as observations.

\[ l(\tilde{\theta}; y) \] the log-likelihood for the saturated model

\[ l(\hat{\theta}; y) \] the log-likelihood of any other model (with MLEs)

The discrepancy of the model under consideration from the saturated model is measured by

\[ S = \frac{D(\hat{\theta}; \hat{\theta})}{\phi} = -2 \left\{ l(\hat{\theta}; y) - l(\hat{\theta}; y) \right\} \]

The numerator of the fraction on the left is called the **deviance** for the model under consideration. The scaled version of this \[ S = \frac{D(\hat{\theta}; \hat{\theta})}{\varphi} \], for some appropriate scale parameter \( \varphi \), is called the **scaled deviance**.

The sampling distribution of \( S \) is asymptotically \( X^2 \) with \( n-p \) degrees of freedom.
Example: The normal distribution

\[ l(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu_i)^2 \]

In the saturated model, the MLE of \( \mu_i \) is just \( y_i \)

\[ l(\tilde{\theta}, y) = -\frac{n}{2} \ln(2\pi\sigma^2) \]

In the current model we have some **linear predictor**:

\[ \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \]

and the fitted values are

\[ \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip} \]

then,

\[ l(\hat{\theta}, y) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]

Scaled deviance (and thus the likelihood-ratio criterion) is

\[ S = \sum_{i=1}^{n} \left( \frac{y_i - \hat{y}_i}{\sigma^2} \right)^2 \]

which have \( X^2_{n-(p+1)} \) distribution if the current model is appropriate. In practice \( \sigma^2 \) is unknown. The deviance \( D=\sigma^2 S \) is the residual sum of squares.
**Nested models**

**Example**

If model 1 has linear predictor for $i$th unit

$$
\mu_i = \beta_0 + \beta_1 x_{i1}
$$

and model 2, with the same distribution, link function and scale parameter, has linear predictor

$$
\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}
$$

then model 1 is nested in model 2 because it can be derived from it by setting $\beta_2=\beta_3=0$.

The relative fit of two nested models can be compared by comparing their respective scaled deviances. The change in scaled deviance comparing models 1 and 2 is

$$
S_1 - S_2 = -2(l_1 - l_s) + 2(l_2 - l_s) = -2(l_1 - l_2)
$$

or minus twice the log likelihood ratio.

If model 1 has $p_1$ parameters and model 2 has $p_2 > p_1$ parameters, then under the null hypothesis that both models fit equally well, and given that each is an adequate fit, then asymptotically $S_1 - S_2 \sim X^2_{p_2-p_1}$.
**Example: Normal distribution**

We have seen that in this case the scaled deviance is \( \text{RSS}/\sigma^2 \), so

\[
S_1 - S_2 = \frac{\text{RSS}_1 - \text{RSS}_2}{\sigma^2}
\]

Under the null hypothesis this has a \( X_{p_2-p_1}^2 \) distribution.

In practice, \( \sigma^2 \) has to be estimated, and the residual mean square from the more complex model is used for this:

\[
\hat{\sigma}^2 = \frac{\text{RSS}_2}{n - p_2}
\]

and the ratio

\[
\frac{(\text{RSS}_1 - \text{RSS}_2)/(p_2 - p_1)}{\hat{\sigma}^2}
\]

is referred to an \( F_{p_2-p_1,n-p_2} \) distribution.
Notes on the deviance

1. The F test for comparing normal linear models is ‘exact’, i.e. not just asymptotic.

2. The $X^2$ test for change in deviance is typically a much better asymptotic approximation than the $X^2$ test for comparison with the saturated model.

3. There is disagreement over the terminology for the deviance. The terms for D and S above are used as the standard reference, McGullagh and Nelder, and in Stata. Clayton and Hills call S the deviance (not scaled deviance).

4. In a Poisson or binomial model, the scale parameter is 1, so S and D coincide and the changes in deviance can be referred directly to a $X^2$ distribution: these tests are log likelihood ratio tests.
Wald tests of hypotheses

Wald tests are based on the asymptotic distribution of the maximum-likelihood estimator $\hat{\theta}$ of $\theta$. Thus, in large samples (under some regularity conditions\(^1\)), the distribution of $\theta$ is

$$\theta \sim N\left(\hat{\theta}, I^{-1}(\theta)\right)$$

a multivariate normal distribution with mean equal to the unknown parameter, and variance-covariance matrix equal to the inverse of the information matrix. The latter is the $n \times n$ matrix of minus the expectation of the second derivatives, i.e.,

$$I(\theta) = -E \left\{ \frac{\partial^2 l}{\partial \theta_i \partial \theta_j} \right\}, \text{ with } i = 1, \ldots, n, j = 1, \ldots, n.$$

\(^1\) Basically these have to do with the existence of the third derivative of the log-likelihood
**Wald tests (continued)**

Using the asymptotic distribution of $\hat{\theta}$, under the null hypothesis $H_0: \theta = \theta_o$, the quadratic form

$$W = \left(\hat{\theta} - \theta_o\right)' \hat{I}(\theta) \left(\hat{\theta} - \theta_o\right) \sim \chi_n^2$$

approximately, where $\hat{I}(\theta)$ is the observed information matrix, i.e., the matrix comprised of minus the second derivatives of the log likelihood without having taken their expectation.

In the univariate case, the square root of the above test

$$z = \frac{\hat{\theta} - \theta_o}{\sqrt{\text{var}(\hat{\theta})}}$$

is asymptotically distributed according to a standard normal distribution. Notice that we do not assume normally distributed $\theta$. In that case, the ratio above would be *exactly* distributed according to a $t$ distribution. Unless this assumption holds, the small-sample properties of the above ratio are unknown.
Example (continued): Binomial distribution.

The observed information in the case of the binomial distribution is

\[
I(\hat{\pi}) = -\frac{d^2 l(\pi; y)}{d\pi^2} = \frac{1}{\pi^2} \sum_{i=1}^{n} y_i - \frac{1}{(1-\pi)^2} \left( \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} n_i \right)\]

\[
= \frac{1}{(0.2)^2} (10) - \frac{1}{(1-0.2)^2} (10 - 50) = 312.5
\]

So given the data presented before, the Wald test for the null hypothesis

\[H_0 : \pi = 0.3\]

would be

\[
\left( \frac{0.2 - 0.3}{\sqrt{1/I(0.2)}} \right) = \frac{-0.10}{\sqrt{1/312.5}} \approx -1.767
\]

which asymptotically is a z test that in the case of a two-sided alternative hypothesis would produce a p-value

\[P(|Z| > 1.767) = 0.077\]
Example: Plasma levels of retinol

Low plasma concentrations of retinol, beta-carotene, or other carotenoids might be associated with increased risk of developing certain types of cancer. The following data (from Therese Stuckel, Dartmouth University), were taken from $n=315$ subjects that had had a surgical procedure during a three-year period that involved biopsy or extraction of non-cancerous lesions. This cross-sectional study explored factors determining the levels of these substances. The data obtained are as follows:

- AGE: Age in years
- SEX: Sex (1=Male, 2=Female).
- SMOKSTAT: Smoking status (1=Never, 2=Former, 3=Current Smoker)
- QUETELET: Quetelet ($\text{weight}/(\text{height}^2)$)
- VITUSE: Vitamin Use (1=Yes, fairly often, 2=Yes, not often, 3=No)
- CALORIES: Number of calories consumed per day.
- FAT: Grams of fat consumed per day.
- FIBER: Grams of fiber consumed per day.
- ALCOHOL: Number of alcoholic drinks consumed per week.
- CHOLESTEROL: Cholesterol consumed (mg per day).
- BETADIET: Dietary beta-carotene consumed (mcg per day).
- RETDIET: Dietary retinol consumed (mcg per day)
- BETAPLASMA: Plasma beta-carotene (ng/ml)
- RETPLASMA: Plasma Retinol (ng/ml)
Example: Plasma levels of retinol

In our example, consider the association of plasma levels of retinol (retplasm, the outcome) with alcohol consumption (alcohol, the predictor). An outlier has been removed from the data.

```
. reg  retplasm alcohol

Source |       SS       df       MS              Number of obs =     314
-------------+------------------------------------------------------
Model |   671843.17     1   671843.17           Prob > F      =  0.0001
    Residual |  12948338.7   312  41501.0855
-------------+------------------------------------------------------
    Total |  13620181.9   313   43514.958           Root MSE      =  203.72

                +--------------------------------/+----------------------------------------
retplasm |      Coef.   Std. Err.      t    P>|t|      [95% Conf. Interval]
-------------+--------------------------------/+----------------------------------------
alcohol |   9.365251   2.327637     4.02   0.000     4.785401     13.9451
_cons     |   578.8857   13.04634    44.37   0.000     553.2158    604.5556
```

Higher alcohol consumption is associated with higher plasma levels of retinol, a possibly carcinogenic substance. The output lists the $F$ test of the overall significance of regression, the R-square and a $t$ test assessing the significance of the predictor (equal here to the square root of the $F$ statistic).
Example (continued)

Contrast this with the output from the STATA command `glm`.

```stata
.glm retplasm alcohol
Iteration 0:   log likelihood = -2113.9991

Generalized linear models                            No. of obs = 314
Optimization : ML: Newton-Raphson                    Residual df = 312
Scale param = 41501.09
Deviance = 12948338.69                               (1/df) Deviance = 41501.09
Pearson = 12948338.69                                (1/df) Pearson = 41501.09

Variance function: V(u) = 1                           [Gaussian]
Link function : g(u) = u                              [Identity]
Standard errors : OIM

Log likelihood = -2113.999055                       AIC = 13.4777
BIC = 12948327.19

+----------------------------------------------------------+
|              | Coef.  | Std. Err. |    z   | P>|z|   |  [95% Conf. Interval] |
|----------------------------------------------------------|
| alcohol       | 9.365251| 2.327637  | 4.02   | 0.000 | 4.803167 13.92734    |
| _cons         | 578.8857| 13.04634  | 44.37  | 0.000 | 553.3153 604.4561   |
+----------------------------------------------------------+
```
The STATA `glm` command

The output above includes the following:

1. Maximization history (number of iterations in the iterative maximization procedure)
2. The value of the maximized log-likelihood (-2113.9991 here)
3. The variance function and link (here 1 and identity respectively)
4. The AIC and BIC numbers (useful for comparing different models)
5. The MLE estimate of $\beta$, which is identical to the OLS estimate presented above
6. A $z$ test of the significance of the predictor (`alcohol`). This is identical to the $t$ test listed in the regression output above (i.e., $\hat{\beta}/\text{var}(\hat{\beta})$), but the distributional assumptions are slightly different in the generalized linear model and the general linear model as we mentioned earlier.
7. The deviance (12948338.69) and the $(1/\text{df})$ Deviance (41501.09), equal respectively to the residual sum of squares and mean squares in the linear model.
Example (continued)

The effect of alcohol consumption on plasma retinol levels is given graphically as follows:

```
. quietly glm retplasm alcohol

. predict yhat
(option mu assumed; predicted mean retplasm)
(1 missing value generated)

. graph yhat retplasm alcohol, xlab ylab c(l.) s(i0) border
```
**Example: Assessing the significance of fat as a predictor of plasma retinol levels**

Consider the following output (note the use of sequential sums of squares – option *sequential*):

```
. anova retplasm alcohol fat, continuous(alcohol fat) seq
```

<table>
<thead>
<tr>
<th>Source</th>
<th>Seq. SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>839986.495</td>
<td>2</td>
<td>419993.248</td>
<td>10.22</td>
<td>0.0001</td>
</tr>
<tr>
<td>alcohol</td>
<td>671843.17</td>
<td>1</td>
<td>671843.17</td>
<td>16.35</td>
<td>0.0001</td>
</tr>
<tr>
<td>fat</td>
<td>168143.325</td>
<td>1</td>
<td>168143.325</td>
<td>4.09</td>
<td>0.0439</td>
</tr>
<tr>
<td>Residual</td>
<td>12780195.4</td>
<td>311</td>
<td>41093.8758</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>13620181.9</td>
<td>313</td>
<td>43514.958</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The sequential option produces $SSE(fat|alcohol)=168143.325$. The $F$ statistic associated with it is $F=4.09$. Comparing this to an $F$ distribution with 1 and 311 degrees of freedom, we obtain a p-value of 0.0439, which is just significant at the 5% alpha level, arguing for addition of fat in the model.
Example (continued): The test command

There are several ways to obtain $SSE(fat|alcohol)=168143.325$ besides using the anova command.

We could for example subtract $SSE(fat,alcohol)=12780195.4$ from $SSE(alcohol)=12948338.69$.

The result would be $SSE(fat|alcohol)=168143.325$.

Alternatively, we can use the test command after having run the reg command as follows:

```
. test fat
( 1)  fat = 0.0

    F(  1,   311) =    4.09
    Prob > F =    0.0439
```

or after the anova command as follows:

```
. test fat

Source | Partial SS    df       MS           F     Prob > F
-------+-------------+-----------+-------------+--------+-----------
        fat | 168143.325     1  168143.325       4.09     0.0439
        Residual | 12780195.4   311  41093.8758

```

Copyright © 2011 Giota Touloumi, Nikos Pantazis, Costantine Yiannoutsos, All rights reserved
Model-building in GLM

Example: Consider the output from the STATA command glm:

```
   . glm  retplasm alcohol fat

   Iteration 0:  log likelihood =  -2111.9469

   Generalized linear models                         No. of obs = 314
   Optimization : ML: Newton-Raphson                Residual df = 311
   Scale param = 41093.88                           (1/df) Deviance = 41093.88
   Deviance = 12780195.36                           (1/df) Pearson = 41093.88
   Pearson = 12780195.36
   Variance function: V(u) = 1                     [Gaussian]
   Link function : g(u) = u                         [Identity]
   Standard errors : OIM

   Log likelihood =  -2111.946946                    AIC = 13.471
   BIC = 12780178.12

   +--------------------------------------------------+
   |              | Coef.  | Std. Err. |    z   |   P>|z|  |       95% Conf. Interval       |
   |              |       |           |        |     |                   |                     |
   +--------------------------------------------------+
   alcohol     |  9.990265 |  2.336708 |  4.28  | 0.000 |  5.410401         | 14.57013            |
   fat         | -0.697524 |  0.344846 | -2.02  | 0.043 | -1.373439         | -0.0216661          |
   _cons       |  630.7705 |  28.7483  | 21.94  | 0.000 |  574.4249         |  687.1162           |
   +--------------------------------------------------+
```
Example (continued):

In the plasma retinol levels example, the likelihood-ratio criterion can be calculated by subtracting the deviance $D(X_1, X_2) = 12780195.36$ from $D(X_1) = 12948338.69$ and dividing by $41093.88$ ($(1/\text{df})$ Deviance). The result is $4.09$, which compared to a chi-square distribution with one degree of freedom produces a $p$ value of $0.0431$ as seen from the following output:

```
. display chi2tail(1,4.09)
.04313765
```

Alternatively, we can use the `test` command as follows:

```
. test fat
   ( 1) [retplasm]fat = 0.0
        chi2(  1) =    4.09
    Prob > chi2 =    0.0431
```
Example: Joint significance of fat and fiber after accounting for alcohol.

To account for the joint significance of daily fat and fiber intake after alcohol consumption has been accounted for we proceed as follows:

```
. reg retplasm alcohol fat fiber
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 314</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>841663.312</td>
<td>3</td>
<td>280554.437</td>
<td>F( 3, 310) = 6.81</td>
</tr>
<tr>
<td>Residual</td>
<td>12778518.5</td>
<td>310</td>
<td>41221.0276</td>
<td>Prob &gt; F = 0.0002</td>
</tr>
<tr>
<td>Total</td>
<td>13620181.9</td>
<td>313</td>
<td>43514.958</td>
<td>R-squared = 0.0618</td>
</tr>
</tbody>
</table>

| retplasm | Coef. | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|-----------|-------|-----------|-------|-----|-----------------|
| alcohol   | 9.964244 | 2.343874  | 4.25  | 0.000 | 5.35233         |
| fiber     | -0.6767318 | 0.3604768 | -1.88 | 0.061 | -1.386023       |
| _cons     | 635.0317 | 35.71259 | 17.78 | 0.000 | 564.762         |
Example (continued)

We could calculate the $F$ criterion $\frac{|SSE(X_1) - SSE(X_1, X_2)|/p_2}{SSE(X_1, X_2)/(n - p_1 - p_2 - 1)}$ manually but the `test` command simplifies things significantly:

```
. test fat fiber
   ( 1)  fat = 0.0
   ( 2)  fiber = 0.0

   F(  2,  310) =   2.06
   Prob > F =   0.1292
```

The result indicates that the two predictors are jointly non-significant at the 95% significance level.

Compare this output to that of the `test` command after fitting a GLM with `alcohol`, `fiber` and `fat` as the predictors of retinol plasma levels (`retplasm`):
. glm retplasm alcohol fat fiber

Iteration 0:  log likelihood = -2111.9263

Generalized linear models                                No. of obs = 314
Optimization     : ML: Newton-Raphson                    Residual df = 310
Scale param     = 41221.03
Deviance         = 12778518.55 (1/df) Deviance = 41221.03
Pearson          = 12778518.55 (1/df) Pearson = 41221.03

Variance function: V(u) = 1                        [Gaussian]
Link function    : g(u) = u                        [Identity]
Standard errors  : OIM

Log likelihood   = -2111.926345                    AIC      = 13.47724
BIC              = 12778495.55

------------------------------------------------------------------------------
|      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
|-------------------------|-------------------------|----------|--------|-------------------------|
| alcohol                |   9.964244   2.343874     4.25   0.000     5.370336    14.55815
| fat                    |  -.6767318   .3604768    -1.88   0.060    -1.383253    .0297898
| fiber                  |  -.4526052   2.244069    -0.20   0.840    -4.850899    3.945688
| _cons                  |   635.0317   35.71259    17.78   0.000     565.0363    705.0271
------------------------------------------------------------------------------

. test fat fiber
( 1)  [retplasm]fat = 0.0
( 2)  [retplasm]fiber = 0.0
    chi2(  2) =    4.12
    Prob > chi2 =    0.1275
Example (continued)

That is,

\[
\frac{|SSE(X_1) - SSE(X_1, X_2)|}{SSE(X_1, X_2)/(n - p_1 - p_2 - 1)} = \frac{169820.142}{12778518.5/310} = \frac{169820.142}{41221.0276} \approx 4.12 = 2 \times (2.06)
\]

Thus, the multiple-regression model and the GLM are producing almost identical results.

Addition of the predictors fat and fiber jointly is not significant in terms of the overall reduction of unexplained variability (error) in the model, after the effect of alcohol has been accounted for.