M.Sc. in Applied Statistics MT2005

Wald, Score and Likelihood Ratio tests

Consider a scalar or vector parameter θ indexing a family of distributions and hence log-likelihood $L(\theta)$. The (log-)likelihood-ratio test (LRT) of $\theta = \theta_0$ against unrestricted $\theta \in \Theta$ is $L(\widehat{\theta}) - L(\theta_0)$ where $\widehat{\theta}$ is the MLE. Under regularity conditions (which include that θ_0 is not a boundary value in Θ), the twice the LRT is asymptotically distributed as χ^2_q, where q is the number of restrictions imposed. (Here θ_0 can either be some fixed value or the MLE in $\Theta_0 \subset \Theta$.)

The LRT requires us to maximize the log-likelihood and also to evaluate it at θ_0. If we wish to consider many possible restrictions (for example to test if many individual parameters are zero) the latter can be inconvenient and avoided by using Wald's approximation, which is to replace the log-likelihood L by a quadratic approximation L_W which agrees with L and its first and second derivatives at $\widehat{\theta}$. Since $\widehat{\theta}$ is the MLE, the first derivative will be zero, and hence

$$L_W(\theta) = L(\widehat{\theta}) + \frac{1}{2}(\theta - \widehat{\theta})^T H(\widehat{\theta})(\theta - \widehat{\theta})$$

where H is the Hessian[1] of L. This is the same log-likelihood we would obtain if we assumed that $\widehat{\theta} \sim N(\theta, J^{-1})$ where $J = -H$ is the *observed information matrix*. Further, we approximate J by $I(\widehat{\theta})$, the Fisher information at the MLE: this is what is most commonly known as the Wald test. (Occasionally $I(\theta_0)$ is used.)

We can use a Wald test either by assuming that $\widehat{\theta} \sim N\left(\theta, I(\widehat{\theta})^{-1}\right)$ or via using one of the approximations as if it were the actual log-likelihood. For example, to test if a single parameter is zero via assuming the t-ratio is $N(0,1)$ is an application of a Wald test.

Alternatively, we might want to consider many extensions to a model, for example adding one out of many new explanatory variables. In that case we may wish to avoid maximizing L under each extended model. This is what Rao's score tests do, by approximating the log-likelihood from its shape at θ_0. The usual form is to consider the score U, the derivative of L with respect to θ at θ_0: if θ_0 were the MLE this would be zero. To see how non-zero it is we compare it to a reference normal distribution, and under the null hypothesis this has covariance matrix the Fisher information $I(\theta_0)$. Thus the score statistic is $U^T I(\theta_0)^{-1} U$.

Score tests can also be derived by approximating L by fitting the quadratic which agrees with L and its first two derivatives at θ_0, that is

$$L_S(\theta) = L(\theta_0) + (\theta - \theta_0)^T L'(\theta_0) + \frac{1}{2}(\theta - \theta_0)^T H(\theta_0))(\theta - \theta_0)$$

and expanding $L'(\theta) \sim L'(\widehat{\theta}) + H(\theta)(\theta - \widehat{\theta})$ if we replace $H(\theta_0)$ by $-I(\theta_0)$.

For regression with known σ^2 the log-likelihood is quadratic and has known curvature (the Hessian is constant in both θ and the data), so all these methods agree exactly. Under standard large-sample theory all these tests have an asymptotic χ^2_q distribution, but in small-sample situations they can differ considerably (and the χ^2_q distribution can be an poor approximation).

[1] the matrix of second derivatives.

As a simple example, consider the Poisson(μ) family, and a test of $\mu = 1$ with a single observation $Y = 4$: see figure 1. (This is an unusually large observation, but is not extreme having probability 1.5%.)

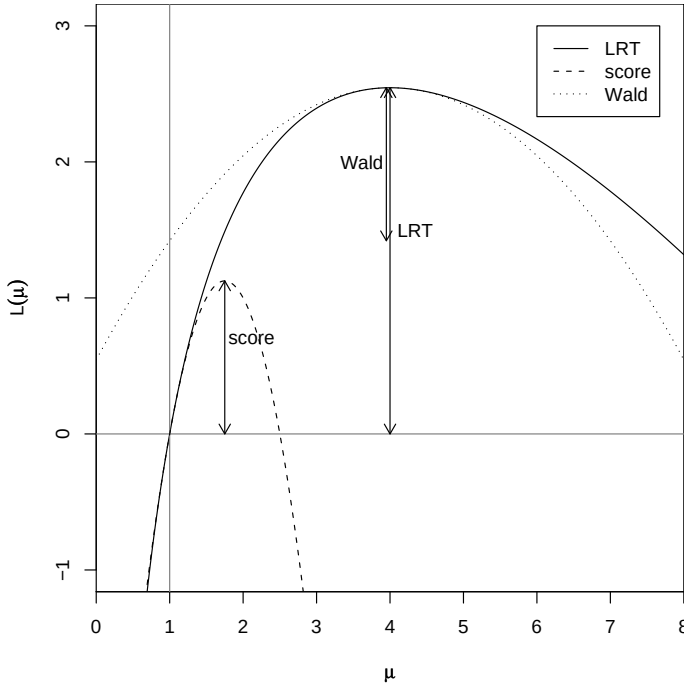

Figure 1: Tests for $\mu = 1$ in a Poisson family with observation $Y = 4$. The exact log-likelihood and the two quadratic approximations are shown.

Note that whereas the LRT is invariant to a change in parametrization, the most of the approximations are not. It is natural to consider log-linear models for a Poisson family, so let us repeat the example for $\log \mu$: see figure 2. In this case the approximations are much better.

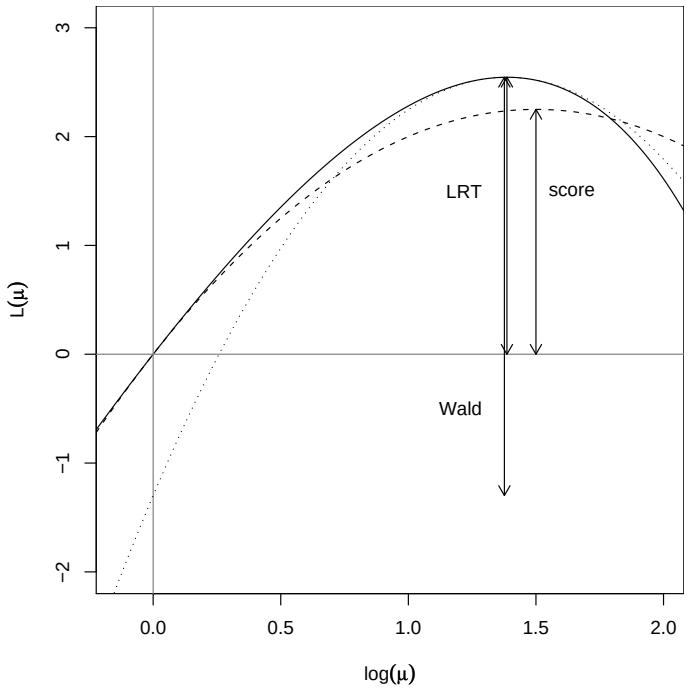

Figure 2: Tests for $\mu = 1$ in a Poisson family with observation $Y = 4$, parameterized by $\theta = \log \mu$. Other details as figure 1.

Deviance and Chi-squared tests

Now suppose we have k Poisson-distributed counts O_i with means μ_i which are functions of the parameter(s) θ. The log-likelihood is

$$L(\theta) = \sum_i O_i \log \mu_i(\theta) - \mu_i(\theta) - \log O_i!$$

and since the saturated model has $\widehat{\mu}_i = O_i$, the deviance is

$$G^2 = 2\sum_i \left[O_i \log\{O_i/\mu_i(\widehat{\theta})\} + (\widehat{\mu}_i(\theta) - O_i)\right]$$

and for most models (including all log-linear model with an intercept)

$$\sum_i (\widehat{\mu}_i(\theta) - O_i) = 0$$

so the G^2 test as normally considered ignores this term. It is conventional to call $\mu(\theta_i)$ the expected value E_i.

Now consider the test of the parametrized means within the general model of freely-specified means. Then

$$L(\mu) = \sum_i O_i \log \mu_i - \mu_i - \log O_i!$$

and so

$$L(\mu) - L(E) = \sum_i O_i \log(\mu_i/E_i)$$

provided $\sum \mu_i = \sum E_i$. The MLE is $\widehat{\mu}_i = O_i$ and the LRT is

$$G^2 = 2L(\widehat{\mu}) - 2L(E) = 2\sum_i O_i \log(O_i/E_i)$$

If we expand this about O_i we get

$$G^2 = -2\sum_i O_i \log(E_i/O_i) \approx 2\sum_i O_i \frac{(O_i - E_i)^2}{2O_i^2} = \sum_i \frac{(O_i - E_i)^2}{O_i}$$

Here the Hessian of L is $-\text{diag}(O_i/\mu_i^2)$ so the Fisher information matrix is $\text{diag}(1/\mu_i)$ which when evaluated at the MLE is the same as the observed information matrix. So the two forms of the Wald approximation agree and give a variation on the Chi-squared test.

For the score test, we first find the first and second derivatives:

$$\frac{\partial L}{\partial \mu_i} = O_i/\mu_i - 1, \qquad \frac{\partial^2 L}{\partial \mu_i \partial \mu_j} = -\delta_{ij} O_i/\mu_i^2, \qquad I(\mu) = \text{diag}(1/\mu_i)$$

Thus the score statistic is

$$U^T I(E)^{-1} U = [O_i/E_i - 1]^T I(E)^{-1} [O_i/E_i - 1] = \sum_i (\frac{O_i}{E_i} - 1)\frac{1}{E_i}(\frac{O_i}{E_i} - 1) = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

the Chi-squared statistic.

The Mantel–Haenszel test

The Mantel–Haenszel test of conditional independence in multiple 2×2 tables is a score test approximation to the LRT in a logistic regression. Suppose we have a $2 \times 2 \times K$ table with counts n_{ijk}. and with variables X, Y and Z, and we are interested in whether

$$\pi_{ik} = P(Y = 2 \,|\, X = i, Z = k)$$

depends on i. The Mantel–Haenszel test statistic is

$$MH = \sum_k \frac{[n_{22k} - \mu_{22k}]^2}{\text{var}(n_{22k})}$$

The full logit model is

$$\text{logit}\, \pi_{ik} = \alpha + \beta x_i + \gamma_k, \text{ for } j = 1, 2 \text{ and } k = 1, \ldots, K$$

where $x_1 = 0$ and $x_2 = 1$, say. The null hypothesis of conditional independence of X and Y given Z corresponds to $\beta = 0$. Writing out the details of the score test (Day and Byar, *Biometrics*, 1979) of $\beta = 0$ leads to MH.

Code

For interest, figure 1 was computed in R (to allow mathematics in the labels) by

```
xx <- seq(0, 8, len = 500)
f <- function(mu) log(dpois(4, mu)/dpois(4,1))
plot(xx, f(xx), type = "l", ylim = c(-1, 3), xaxs = "i",
     xlab = expression(mu), ylab = expression(L(mu)) )
abline(h = 0, col = "gray50")
abline(v = 1, col = "gray50")
arrows(4, 0, 4, f(4), length = 0.1, code = 3)
text(4.1, 1.5, "LRT", adj = 0)

fW <- function(mu) f(4) - (mu-4)^2/8
lines(xx, fW(xx), lty = 3)
arrows(3.95, fW(1), 3.95, f(4), length = 0.1, code = 3)
text(3.9, 2, "Wald", adj = 1)

fS <- function(mu) 3*(mu-1) - 2*(mu-1)^2
lines(xx, fS(xx), lty = 2)
arrows(1.75, 0, 1.75, fS(1.75), length = 0.1, code = 3)
text(1.8, 0.7, "score", adj = 0)

legend(6, 3, c("LRT", "score", "Wald"), lty = 1:3)
```