STAT 350: Lecture 33

Theory of Generalized Linear Models

If Y has a Poisson distribution with parameter μ then

$$P(Y = y) = \frac{\mu^y e^{-\mu}}{y!}$$

for y a non-negative integer. We can use the method of maximum likelihood to estimate μ if we have a sample $Y_1, \ldots, Y_n$ of independent Poisson random variables all with mean μ. If we observe $Y_1 = y_1$, $Y_2 = y_2$ and so on then the likelihood function is

$$P(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_{i=1}^{n} \frac{\mu^{y_i} e^{-\mu}}{y_i!} = \frac{\mu^{\sum y_i} e^{-n\mu}}{\prod y_i!}$$

This function of μ can be maximized by maximizing its logarithm, the log likelihood function. We set the derivative of the log likelihood with respect to μ equal to 0, getting

$$\frac{d}{d\mu}[\sum y_i \log \mu - n\mu - \sum \log(y_i!)] = \sum y_i/\mu - n = 0\,.$$

This is the **likelihood equation** and its solution $\hat{\mu} = \bar{y}$ is the **maximum likelihood** estimate of μ.

In a regression problem all the Y_i will have different means μ_i. Our log-likelihood is now

$$\sum y_i \log \mu_i - \sum \mu_i - \sum \log(y_i!)$$

If we treat all n of the μ_i as unknown parameters we can maximize the log likelihood by setting each of the n partial derivatives with respect to μ_k for k from 1 to n equal to 0. The kth of these n equations is just

$$y_k/\mu_k - 1 = 0\,.$$

This leads to $\hat{\mu}_k = y_k$. In glm jargon this model is the **saturated** model.

A more useful model is one in which there are fewer parameters but more than 1. A typical glm model is

$$\mu_i = \exp(x_i^T \beta)$$

where the x_i are covariate values for the ith observation (often including an intercept term just as in standard linear regression).

In this case the log-likelihood is

$$\sum y_i x_i^T \beta - \sum \exp(x_i^T \beta) - \sum \log(y_i!)$$

which should be treated as a function of β and maximized.

The derivative of this log-likelihood with respect to β_k is

$$\sum y_i x_{ik} - \sum \exp(x_i^T \beta) x_{i,k} = \sum (y_i - \mu_i) x_{i,k}$$

If β has p components then setting these p derivatives equal to 0 gives the **likelihood equations**.

It is no longer possible to solve the likelihood equations analytically. We have, instead, to settle for numerical techniques. One common technique is called **iteratively re-weighted least squares**. For a Poisson variable with mean μ_i the variance is $\sigma_i^2 = \mu_i$. Ignore for a moment the fact that if we knew σ_i we would know μ_i and consider fitting our model by least squares with the σ_i^2 known. We would minimize (see our discussion of weighted least squares)

$$\sum \frac{(Y_i - \mu_i)^2}{\sigma_i^2}$$

by taking the derivative with respect to β_k and (again ignoring the fact that σ_i^2 depends on β_k we would get

$$-2 \sum \frac{(Y_i - \mu_i) \partial \mu_i / \partial \beta_k}{\sigma_i^2} = 0$$

But the derivative of μ_i with respect to β_k is $\mu_i x_{ik}$ and replacing σ_i^2 by μ_i we get the equation

$$\sum (Y_i - \mu_i) x_{ik} = 0$$

exactly as before. This motivates the following estimation scheme.

1. Begin with a guess for the standard deviations σ_i (taking them all equal to 1 is simple).

2. Do (non-linear) weighted least squares using the guessed weights. Get estimated regression parameters $\hat{\beta}$.

3. Use these to compute estimated variances $\hat{\sigma}_i^2$. Go back to do weighted least squares with these weights.

4. Iterate (repeat over and over) until estimates not really changing.

NOTE: if the estimation converges then the final estimate is a **fixed point** of the algorithm which solves the equation

$$\sum(Y_i - \mu_i)x_{ik} = 0$$

derived above.

Non-linear least squares

In a regression model of the form

$$Y_i = \mu_i + \epsilon_i$$

we call the model linear if $\mu_i = x_i^T\beta$ for some covariates x_i and a parameter β. Often, however, we have **non-linear** models such as $\mu_i = \beta_0(1-\exp(-\beta_1 x_i))$ or in general $\mu_i = f(x_i, \beta)$ for some function f which need not be linear in β. If the errors are homo-scedastic normal errors then the maximum likelihood estimates of the β_i are least squares estimates, minimizing $\sum(Y_i - \mu_i)^2$. The value $\hat{\beta}$ then solves

$$\sum(Y_i - \mu_i)\frac{\partial\mu_i}{\partial\beta} = 0\,.$$

This equation can usually not be solved analytically. In general the process of solving it numerically is iterative. One approach is:

1. Begin with an **initial guess** (no general method available to do this step) for the parameter estimates, say $\hat{\beta}^{(0)}$.

2. **Linearize** the function $\mu_i(\beta)$ around $\hat{\beta}^{(0)}$ by doing a Taylor expansion. This involves writing

$$\mu_i(\beta) \approx \mu_i(\hat{\beta}^{(0)}) + x_i^T(\beta - \hat{\beta}^{(0)})$$

where x_i is the vector of partial derivatives of μ_i with respect to the entries in β (computed using $\hat{\beta}^{(0)}$).

3. Replace μ_i in the sum of squares by this approximation and get that you are trying to minimize
$$\sum (Y_i - \mu_i(\hat\beta^{(0)}) - x_i^T(\beta - \hat\beta^{(0)}))^2$$
with respect to β.

4. This is really now an ordinary least squares problem in which we are regressing $Y_i^* = Y_i - \mu_i(\hat\beta^{(0)})$ on x_i^T. Let X_0 be the $n \times p$ matrix with ith row $x_i^T = \partial\mu_i/\partial\beta$ where the derivative is evaluated at $\hat\beta^{(0)}$. Then we are to minimize
$$|Y^* - X_0\delta|^2$$
where $\delta = \beta - \hat\beta^{(0)}$. The solution is $\hat\delta = (X_0^T X_0)^{-1} X_0^T Y^*$. We compute this δ.

5. Update $\hat\beta^{(0)}$ to $\hat\beta^{(1)} = \hat\beta^{(0)} + \hat\delta$.

6. Recompute the entries in X with the new value $\hat\beta^{(1)}$, get a new $\hat\delta$ then a $\hat\beta^{(2)}$ and so on.

7. Continue until the estimates change little or the sum of squares changes little.

Example

In thermoluminescence dating a common model is
$$Y_i = \beta_1(1 - \exp(-(d_i - \beta_2)/\beta_3)) + \epsilon_i$$
Thus in this case
$$\mu_i(\beta) = \beta_1(1 - \exp(-(d_i - \beta_2)/\beta_3)$$
The derivatives of μ_i with respect to the entries of β are
$$\frac{\partial\mu_i}{\partial\beta_1} = (1 - \exp(-(d_i - \beta_2)/\beta_3)$$
$$\frac{\partial\mu_i}{\partial\beta_2} = -\beta_1 \exp(-(d_i - \beta_2)/\beta_3)/\beta_3$$
and
$$\frac{\partial\mu_i}{\partial\beta_3} = -\beta_1(d_i - \beta_2)\exp(-(d_i - \beta_2)/\beta_3)/\beta_3^2$$

Software and computing details

In SAS the procedure `proc nlin` will do non-linear least squares. You need to specify quite a few ingredients. The `model` statement must specify the exact functional form of the function $\mu_i(\beta)$. You have a statement `parms` which is needed to specify the initial guess for β. You usually will have a bunch of `der` statements to specify derivatives of the predicted values with respect to each parameter and, if you use some algorithms, you will even specify some second derivatives. (You get a choice of 5 different iterative algorithms. The one described above is called **Gauss-Newton**.)

In Splus the function `nls` can be used to to the same thing.