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.

- Set derivative of log likelihood with respect to μ equal to 0.
- Get **likelihood equation**:

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

- 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 include 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 guess for SDs σ_i (taking all to be 1 is easy).
  2. Do (non-linear) weighted least squares using 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 stop 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.

Estimating equations: an introduction via glim

Get estimates $\hat{\theta}$ by solving $h(X, \theta) = 0$ for θ.

1. The normal equations in linear regression:

$$X^T Y - X^T X \beta = 0$$

2. Likelihood equations; if $\ell(\theta)$ is log-likelihood:

$$\frac{\partial \ell}{\partial \theta} = 0.$$

3. Non-linear least squares:

$$\sum (Y_i - \mu_i) \frac{\partial \mu_i}{\partial \theta} = 0$$

4. The iteratively reweighted least squares estimating equation:

$$\sum \frac{Y_i - \mu_i}{\sigma_i^2} \frac{\partial \mu_i}{\partial \theta} = 0;$$

for generalized linear model σ_i^2 is *known* function of μ_i.

Poisson regression revisited

- The likelihood function for a Poisson regression model is:

$$L(\beta) = \prod \frac{\mu_i^{y_i}}{y_i!} \exp(-\sum \mu_i)$$

- the log-likelihood is

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

- A typical glm model is

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

  where x_i is covariate vector for observation i (often include intercept term 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**.
- For a Poisson model the variance is given by

$$\sigma_i^2 = \mu_i = \exp(x_i^T \beta)$$

- so the likelihood equations can be written as

$$\sum \frac{(y_i - \mu_i) x_{i,k} \mu_i}{\mu_i} = \sum \frac{(y_i - \mu_i)}{\sigma_i^2} \frac{\partial \mu_i}{\partial \beta_k} = 0$$

which is the fourth equation above.

IRWLS

- Equations solved iteratively, as in non-linear regression, but iteration now involves *weighted* least squares.
- Resulting scheme is called **iteratively reweighted least squares**.
  1. Begin with guess for SDs σ_i (taking all equal to 1 is simple).
  2. Do (non-linear) weighted least squares using the guessed weights. Get estimated regression parameters $\hat{\beta}^{(0)}$.
  3. Use to compute estimated variances $\hat{\sigma}_i^2$. Re-do weighted least squares with these weights; get $\hat{\beta}^{(1)}$.
  4. Iterate (repeat over and over) until estimates not really changing.

Fixed Points of Algorithms

- Suppose the $\hat\beta^{(k)}$ converge as $k \to \infty$ to something, say, $\hat\beta$.
- Recall

$$\sum \left[\frac{y_i - \mu_i(\hat\beta^{(k+1)})}{\sigma_i^2(\hat\beta^{(k)})}\right] \frac{\partial \mu_i(\hat\beta^{(k+1)})}{\partial \hat\beta^{(k+1)}} = 0$$

- we learn that $\hat\beta$ must be a root of the equation

$$\sum \left[\frac{y_i - \mu_i(\hat\beta)}{\sigma_i^2(\hat\beta)}\right] \frac{\partial \mu_i(\hat\beta)}{\partial \hat\beta} = 0$$

which is the last of our example estimating equations.

Distribution of Estimators

- **Distribution Theory**: compute distribution of statistics, estimators and pivots.
- Examples: Multivariate Normal Distribution; theorems about chi-squared distribution of quadratic forms; theorems that F statistics have F distributions when null hypothesis true; theorems that show a t pivot has a t distribution.
- **Exact Distribution Theory**: exact results as in previous example when errors are assumed to have *exactly* normal distributions.
- **Asymptotic** or **Large Sample Distribution Theory**: same sort of conclusions but only approximately true and assuming n is large. Theorems of the form:

$$\lim_{n\to\infty} P(T_n \le t) = F(t)$$

- For generalized linear models do asymptotic distribution theory.

Uses of Asymptotic Theory: principles

- An estimate is normally only useful if it is equipped with a measure of uncertainty such as a standard error.
- A standard error is a useful measure of uncertainty provided the error of estimation $\hat\theta - \theta$ has approximately a normal distribution and the standard error is the standard deviation of this normal distribution.
- For many estimating equations $h(Y, \theta) = 0$ the root $\hat\theta$ is unique and has the desired approximate normal distribution, **provided the sample size** n **is large**.

Sketch of reasoning in special case

- Poisson example: $p = 1$
- Assume Y_i has a Poisson distribution with mean $\mu_i = e^{x_i \beta}$ where now β is a scalar.
- The estimating equation (the likelihood equation) is

$$U(\beta) = h(Y_1, \ldots, Y_n, \beta) = \sum (Y_i - e^{x_i \beta}) x_i = 0$$

- It is now important to distinguish between a value of β which we are trying out in the estimating equation and the true value of β which I will call β_0.
- If we happen to try out the true value of β in U then we find

$$E_{\beta_0}(U(\beta_0)) = \sum x_i E_{\beta_0}(Y_i - \mu_i) = 0$$

- On the other hand if we try out a value of β other than the correct one we find

$$E_{\beta_0}(U(\beta)) = \sum x_i(e^{x_i\beta} - e^{x_i\beta_0}) \neq 0\,.$$

- But $U(\beta)$ is a sum of independent random variables so by the law of large numbers (law of averages) must be close to its expected value.
- This means: if we stick in a value of β far from the right value we will not get 0 while if we stick in a value of β close to the right answer we will get something close to 0.
- This can sometimes be turned in to the assertion:
  The glm estimate of β is **consistent**, that is, it converges to the correct answer as the sample size goes to ∞.

- The next theoretical step is another **linearization**.
- If $\hat\beta$ is the root of the equation, that is, $U(\hat\beta) = 0$, then

$$0 = U(\hat\beta) \approx U(\beta_0) + (\hat\beta - \beta_0)U'(\beta_0)$$

- This is a **Taylor's expansion**.
- In our case the derivative U' is

$$U'(\beta) = -\sum x_i^2 e^{x_i\beta}$$

so that approximately

$$\hat\beta = \frac{\sum (Y_i - \mu_i)x_i}{\sum x_i^2 e^{x_i\beta_0}}$$

- The right hand side of this formula has expected value 0, variance

$$\frac{\sum x_i^2 Var(Y_i)}{\left(\sum x_i^2 e^{x_i\beta_0}\right)^2}$$

which simplifies to

$$\frac{1}{\sum x_i^2 e^{x_i\beta_0}}$$

- This means that an approximate standard error of $\hat\beta$ is

$$\frac{1}{\sqrt{\sum x_i^2 e^{x_i\beta_0}}}$$

that an estimated approximate standard error is

$$\frac{1}{\sqrt{\sum x_i^2 e^{x_i\hat\beta}}}$$

- Finally, since the formula shows that $\hat\beta - \beta_0$ is a sum of independent terms the central limit theorem suggests that $\hat\beta$ has an approximate normal distribution and that

$$\sqrt{\sum x_i^2 e^{x_i\hat\beta}}(\hat\beta - \beta_0)$$

is an approximate pivot with approximately a $N(0,1)$ distribution.
- You should be able to turn this assertion into a 95% (approximate) confidence interval for β_0.

Scope of these ideas

The ideas in the above calculation can be used in many contexts.

- We can get approximate standard errors in non-linear regression.
- We can get approximate standard errors in any model where we do maximum likelihood.
- We can show that the assumption of normal errors does not have too big an impact on the t and F tests in multiple regression.
- We can get approximate standard errors in generalized linear models.
- We can demonstrate that the role of the Error Sum of Squares in multiple regression can be replaced, approximately, by a function called the **Deviance** which is a function whose *derivative* (with respect to the parameters) is the estimating equation.