

## EM Algorithm for Estimating Equations

George Casella\* and Pamela A. Ohman †

June 5, 1998

**Keywords:** consistency**Abstract****1 Introduction****2 The Algorithm**

The EM algorithm can be generalized as follows:

Assume a density for the missing data given the observed is  $f(z|y; \theta)$ . Further, suppose that if the full data were known, the values of  $\theta$  that maximize the function  $O_C(y, z; \theta)$  would provide estimates for that parameter. Thus, call  $O_C(y, z; \theta)$  the complete objective function. Let  $O_I(y; \theta)$  represent the incomplete objective function obtained by integrating  $O_C(y, z; \theta)$  with respect to the density  $f(z|y; \theta)$ . That is,

$$O_I(y; \theta) = \int_{\mathcal{Z}} O_C(y, z; \theta) f(z|y; \theta) dz$$

Assume

---

\*Cornell University, Supported by National Science Foundation Grant DMS-9305547

†University of Florida

1.  $0 < O_C(y, z; \theta) < \infty$ .

2.  $O_I(y; \theta) < \infty$  for all  $\theta$ ,

Define a modified conditional density for the missing data as

$$k(z|y; \theta) = \frac{O_C(y, z; \theta)f(z|y; \theta)}{O_I(y; \theta)},$$

so that

$$\log [O_C(y, z; \theta)f(z|y; \theta)] = \log O_I(y; \theta) + \log k(z|y; \theta).$$

Our goal will be to maximize the quantity  $O_I(y; \theta)$  with respect to  $\theta$ , or alternatively (if equivalent) to solve  $\frac{d}{d\theta} O_I(y; \theta) = 0$ . This can be done by iterating between the following steps, starting with an initial value of  $\theta^{(0)}$ , so that at the end of the  $t$ 'th iteration one has obtained the estimate  $\theta^{(t)}$ .

- **E-step** Average out  $z$  with respect to the weighting function  $k(z|y; \theta^{(t)})$ .

$$Q(\theta|\theta^{(t)}) = \log O_I(y; \theta) + H(\theta|\theta^{(t)}),$$

where  $Q(\theta|\theta^{(t)}) = \int \log [O_C(y, z; \theta)f(z|y; \theta)] k(z|y; \theta^{(t)}) dz$  and  $H(\theta|\theta^{(t)}) = \int \log k(z|y; \theta) k(z|y; \theta^{(t)}) dz$ .

- **M-step** Maximize  $Q(\theta|\theta^{(t)})$  with respect to  $\theta$ . (In some cases, it may be more convenient to solve the equation  $\frac{d}{d\theta} Q(\theta|\theta^{(t)}) = 0$  for  $\theta$ .)

The net result of these two steps is that we may write a sequence of estimates as

$$\theta^{(t+1)} = \operatorname{argmax}_{\theta} \int \log [O_C(y, z; \theta)f(z|y; \theta)] k(z|y; \theta^{(t)}) dz$$

Note that the above algorithm works because  $k(z|y, \theta)$  behaves like a density. It is greater than or equal to zero and it integrates to 1. In particular, for each iteration,  $O_I(y; \theta^{(t)})$  is greater than at the previous iteration. (MORE PROOF OF ACTUAL CONVERGENCE IS NEEDED!!) That is

$$\log O_I(y; \theta^{(t+1)}) - \log O_I(y; \theta^{(t)}) \geq 0.$$

This follows because

$$Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) \geq 0$$

by construction. Furthermore,

$$\begin{aligned} H(\theta^{(t)}|\theta^{(t)}) - H(\theta^{(t+1)}|\theta^{(t)}) &= \int \log \left[ \frac{k(z|y; \theta^{(t)})}{k(z|y; \theta^{(t+1)})} \right] \times \left[ \frac{k(z|y; \theta^{(t)})}{k(z|y; \theta^{(t+1)})} \right] k(z|y; \theta^{(t+1)}) dz \\ &\geq \log \int \frac{k(z|y; \theta^{(t)})}{k(z|y; \theta^{(t+1)})} k(z|y; \theta^{(t+1)}) dz \times \int \frac{k(z|y; \theta^{(t)})}{k(z|y; \theta^{(t+1)})} k(z|y; \theta^{(t+1)}) dz \\ &= \log \int k(z|y; \theta^{(t)}) dz \times \int k(z|y; \theta^{(t)}) dz \\ &= \log 1 \times 1 \\ &= 0 \end{aligned}$$

The inequality above follows from the fact that  $x \log x$  is a convex function and  $G(y|z; \theta)$  can be thought of as a density.

We possibly also need to specify that

- $\int \log k(y, z; \theta) k(z|y; \theta^{(t)}) dz < \infty$
- $\int \log k(y|z; \theta) k(z|y; \theta^{(t)}) dz < \infty$

I think the latter is automatic since  $k(z|y; \theta^{(t)})$  behaves like a density. The former may be automatic as well since  $k(z|y; \theta)$  is a density-like function.

## 2.1 Implementation for Estimating Equations

We would like to apply this algorithm to finding roots of estimating equations. In all cases, it will be necessary to find a quantity, corresponding to the estimating equation of interest, such that maximizing that quantity is equivalent to solving the estimating equation of interest such that conditions 1-2 are satisfied.

The EM algorithm will maximize the quantity  $O_I(y; \theta)$  with respect to the unknown parameter  $\theta$ . Under standard regularity conditions, this is equivalent to finding the roots of the function  $g(y; \theta) = \frac{d}{d\theta} O_I(y; \theta)$ . Thus, again under some standard conditions, we can conclude that if  $g(y; \theta)$  is an unbiased estimating equation then its roots, or equivalently the values that maximize  $O_I(y; \theta)$ , are consistent estimates of  $\theta$ . Alternative conditions for consistency are discussed in section 4 as well.

## 3 Special Cases

### 3.1 Least Squares for Regression with Missing Covariates

For this example, assume a simple linear regression, where  $y$  is the vector of independently distributed response variables of length  $n$  and  $x$ , the corresponding vector of covariates. The variables are related through the linear equation

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Further, suppose that we can write  $x = (x'_o, x'_m)'$ , such that  $x_o$  is the vector of observed covariates and  $x_m$  is the vector of missing covariates ( $z$ ). The usual least squares estimating functions, given that all of the data are observed, are

$$\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) \tag{1}$$

and

$$\sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i). \quad (2)$$

The objective function which is to be minimized is the sum of squares

$$\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

This may be converted into an appropriate objective function which is to be maximized. In particular, define

$$O_C(y, x_o, x_m; \beta) = \exp \left\{ - \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \right\}.$$

Maximizing this new objective function results in the same estimates as obtained by finding the roots of the original least squares estimating equations (1) and (2).

We now choose  $f(x_m|x_o, \beta)$  as the density of the missing data given the observed. Note that we let each of the missing covariates be independently distributed as normal random variables with mean  $\mu(y_i, \beta)$  and some variance  $\sigma_x^2$ .

The incomplete objective function is then defined as

$$\begin{aligned} O_I(y, x_o; \beta) &= \int O_C(y, x_o, x_m; \beta) f(x_m|x_o, y; \beta) dx_m \\ &= \int \exp \left\{ - \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \right\} f(x_m|x_o, y; \beta) dx_m \end{aligned}$$

Further,

$$k(x_m|y, x_o; \beta) = \frac{O_C(y, x_o, x_m; \beta) \times f(x_m|x_o, y; \beta)}{\int O_C(y, x_o, x_m; \beta) \times f(x_m|x_o, y; \beta) dx_m}$$

The quantity to be minimized is

$$O_I(y, x_o; \beta) = \int \exp \left\{ - \sum (y_i - \beta_0 - \beta_1 x_i)^2 \right\} f(x_m|y, x_o; \beta) dx_m.$$

However, the quantity that we will be working with to implement the algorithm is

$$Q(\beta|\beta^{(t)}) = \int \left\{ - \sum (y_i - \beta_0 - \beta_1 x_i)^2 + \log f(x_m|x_o, y; \beta) \right\} k(x_m|y, x_o; \beta^{(t)}) dx_m,$$

where

$$k(x_m|y, x_o; \beta) dx_m = \frac{\exp\{-\sum(y_i - \beta_0 - \beta_1 x_i)^2\} f(x_m|y, x_o, \beta)}{\int \exp\{-\sum(y_i - \beta_0 - \beta_1 x_i)^2\} f(x_m|y, x_o, \beta) dx_m}$$

If we assume the maximum is attained by setting the derivative equal to zero, and that the integral and derivative of the resulting equation are interchangeable, then maximizing the above with respect to  $\beta_0$  and  $\beta_1$  will be equivalent to solving the equations

$$\int \left[ 2 \sum (y_i - \beta_0 - \beta_1 x_i) + \frac{d}{d\beta_0} \log f(x_m|y, x_o; \beta) \right] \times \exp \left\{ - \sum (y_i - \beta_0^{(t)} - \beta_1^{(t)} x_i)^2 \right\} f(x_m|y, x_o; \beta^{(t)}) dx_m = 0 \quad (3)$$

and

$$\int \left[ 2 \sum x_i (y_i - \beta_0 - \beta_1 x_i) + \frac{d}{d\beta_1} \log f(x_m|y, x_o; \beta) \right] \times \exp \left\{ - \sum (y_i - \beta_0^{(t)} - \beta_1^{(t)} x_i)^2 \right\} f(x_m|y, x_o; \beta^{(t)}) dx_m = 0 \quad (4)$$

Further, assume that the  $x_i$ 's are independently distributed as normal random variables with mean  $\mu_x$  and variance  $\sigma_x^2$ , which do not depend on the parameters of interest. For implementation of the algorithm in this case, observe that the forms of the density  $f(x_m)$  and  $O_I(y, x_o, x_m; \beta)$  now allow simplification of the forms of (3) and (4), the working estimating equations. To demonstrate, suppose that only one covariate,  $x_m$ , is missing. The first estimating equation can be written

$$\begin{aligned} \frac{1}{2}(3) &= \int \sum (y_i - \beta_0 - \beta_1 x_i) \exp \left\{ - \sum (y_i - \beta_0^{(t)} - \beta_1^{(t)} x_i)^2 \right\} f(x_m) dx_m \\ &= \sum_{j \neq m} (y_j - \beta_0 - \beta_1 x_j) \exp \left\{ - \sum_{j \neq m} (y_j - \beta_0 - \beta_1 x_j)^2 \right\} E_{x_m} \left[ \exp \left\{ - (y_m - \beta_0 - \beta_1 x_m)^2 \right\} \right] \\ &\quad + \exp \left\{ - \sum_{j \neq m} (y_j - \beta_0 - \beta_1 x_j)^2 \right\} E_{x_m} \left[ (y_m - \beta_0 - \beta_1 x_m) \exp \left\{ - (y_m - \beta_0 - \beta_1 x_m)^2 \right\} \right]. \end{aligned}$$

Setting this equal to zero, we further reduce this to

$$\begin{aligned} 0 &= \sum_{j \neq m} (y_j - \beta_0 - \beta_1 x_j) E_{x_m} \left[ \exp \left\{ - (y_m - \beta_0 - \beta_1 x_m)^2 \right\} \right] \\ &\quad + E_{x_m} \left[ (y_m - \beta_0 - \beta_1 x_m) \exp \left\{ - (y_m - \beta_0 - \beta_1 x_m)^2 \right\} \right]. \end{aligned}$$

Through writing out the integral and completing the squares, we get

$$0 = \exp\{f(c)\} \left( \sum_{j \neq m} (y_j - \beta_0 - \beta_1 x_j) + E_{x_m} [y_m - \beta_0 - \beta_1 x_m^*] \right), \quad (5)$$

where

$$x_m^* \sim N(\mu^*, \sigma^{*2})$$

such that

$$\mu^* = \frac{2}{1 + 2\sigma_x^2 \beta_1^{(t)2}} \left( \frac{c}{2} + \beta_1^{(t)} \sigma_x^2 (y_m - \beta_0^{(t)}) \right)$$

and

$$\sigma_x^{*2} = \frac{\sigma_x^2}{1 + 2\sigma_x^2 \beta_1^{(t)2}}.$$

Further,

$$f(c) = \frac{-2\sigma_x^*}{1 + 2\sigma_x^2 \beta_1^{(t)2}} \left( \frac{c}{2\sigma_x^2} + \beta_1^{(t)} (y_m - \beta_0^{(t)}) \right)^2 + \frac{c^2}{2\sigma_x^2} + \beta_1^{(t)} (y_m - \beta_0^{(t)})^2.$$

Thus, from (5),

$$0 = \sum_{j \neq m} (y_j - \beta_0 - \beta_1 x_j) + (y_m - \beta_0 - \beta_1 \mu^*).$$

Likewise, in considering the complementary working estimating equation (the derivative of  $Q(\beta; \beta^{(t)})$  with respect to  $\beta_1$ ),

$$\begin{aligned} \frac{1}{2}(4) &= \int \sum x_i (y_i - \beta_0 - \beta_1 x_i) \exp \left\{ - \sum (y_i - \beta_0^{(t)} - \beta_1^{(t)} x_i)^2 \right\} f(x_m) dx_m \\ &= \sum_{j \neq m} x_j (y_j - \beta_0 - \beta_1 x_j) + (y_m - \beta_0) \mu^* - \beta_1 (\sigma_x^{*2} + \mu^{*2}) \\ &= 0. \end{aligned}$$

Although, we have only demonstrated this for a single missing covariate, the calculations show that the natural extension for multiple missing covariates holds. Thus the estimating equations used at each iteration of the EM algorithm are essentially the estimating equations of the full data, substituting the values of the missing covariates by their expected values. In addition, similar

simplification of the working estimating equations would also be possible if the  $x_i$ 's were independent exponential random variables.

### 3.2 Non-linear regression

The ideas above carry over directly to the case such that the mean is not linear in the covariates.

Assume that

$$E[Y_i|X_i = x_i] = \mu(\beta, x_i).$$

In general, under the assumption of constant variance and independence, we may use

*stuff*

as the objective function, such that the function which we will be using within the EM algorithm is

$$Q(\beta|\beta^{(t)}) = \text{stuff}.$$

Further, the problem of maximizing this can be reduced to that of solving

*and more stuff*

for  $\beta$ .

### 3.3 Quasi-likelihood

Consider that we can in general write

$$E(y_i) = \mu(x_i, \beta).$$

Assume that the observations  $y_i$  are independent. Then we may consider estimating functions of the form

$$\sum_{i=1}^n a_i(\beta) (y_i - \mu(x_i, \beta)).$$

Results by ????? suggest that “optimal” estimating functions of this form are given by

$$\frac{d\mu_i(x_i, \beta)}{d\beta} \sum_{i=1}^n \frac{y_i - \mu_i(x_i, \beta)}{\sigma^2 V_i(\mu_i(x_i, \beta))}.$$

Note that the quasi-likelihood function that yields these estimating equations is

$$\sum_{i=1}^n \int_{y_i}^{\mu_i(x_i, \beta)} \frac{(y_i - t)}{\sigma^2 V_i(t)} dt.$$

This negative function is maximized in order to find estimates for the parameters  $\beta$ . It is not bounded below by zero. To resolve this so that condition (2) is satisfied, let the complete objective function be

$$O_C(y, x; \beta) = \exp \left\{ \sum_{i=1}^n \int_{y_i}^{\mu_i(x_i, \beta)} \frac{(y_i - t)}{\sigma^2 V_i(t)} dt \right\}$$

Thus, the function that will be used in the EM algorithm is given by ...

## 4 Consistency

In both of the cases introduced above, we wish to demonstrate that our estimates are consistent. One way to establish consistency is to demonstrate that the incomplete estimating equations are unbiased. To do this, we present the following sufficient conditions on the complete estimating equations and the density of the missing data.

**Theorem 1** *Let the density of  $z$  given  $y$  be independent of the parameter  $\theta$ . If the complete estimating equation  $g_C(y, z; \theta)$  is unbiased, then the incomplete estimating equation, defined by*

$$g_I(y; \theta) = \int g_C(y, z; \theta) f(z|y) dz$$

*will also be unbiased.*

**PROOF:**

$$E[g_I(y; \theta)] = \int_Y g_I(y; \theta) f(y; \theta) dy$$

$$\begin{aligned}
&= \int_Y \left[ \int_Z g_C(y, z; \theta) f(z|y) dz \right] f(y; \theta) dy \\
&= \int_Y \int_Z g_C(y, z; \theta) f(y, z; \theta) dz dy \\
&= E[g_C(y, z; \theta)] \\
&= 0
\end{aligned}$$

■

This theorem can be applied directly to show that the estimating equations for linear regression presented in the previous section are unbiased when the  $y_i$ 's are symmetrically distributed about their means.

Of interest is whether

$$E\left(\frac{d}{d\beta} O_I(y, x_o; \beta)\right) = 0. \quad (6)$$

**Theorem 2** *Assume that given  $x_i$ , the expectation of  $t_i e^{-t_i^2}$  is zero, where  $t_i = y_i - \mu(\beta, x_i)$ . In addition, assume that the distribution of the missing data  $x_m$  does not depend on  $\beta$ , or  $y_m$ . Then, the incomplete estimating equation,  $\frac{d}{d\beta} O_I(y, x_o; \beta)$ , is unbiased, where*

$$O_I(y, x_o; \beta) = \int \exp\left\{-\sum (y_i - \mu(\beta, x_i))^2\right\} f(x_m) dx_m.$$

**PROOF:** Since the mean and variance of  $x_m$  do not depend on the parameters or observed data, we can simply write  $f(x_m|y, x_o, \beta) = f(x_m)$ . According to the previous theorem, it suffices to consider the expectations of the complete data estimating function

$$E\left[-2\frac{d\mu(\beta, x_i)}{d\beta} \sum_{i=1}^n (y_i - \mu(\beta, x_i)) \exp\left\{-\sum (y_i - \mu(\beta, x_i))^2\right\}\right]. \quad (7)$$

Regardless of the number of parameters in the vector  $\beta$ , the value of this expectation is determined by the value of

$$E\left[\sum_{i=1}^n (y_i - \mu(\beta, x_i)) \exp\left\{-\sum (y_i - \mu(\beta, x_i))^2\right\}\right].$$

Start by looking at the iterated expectation, conditioning first on the full set of covariates,  $x$ .

Letting  $t_i = \mu(\beta, x_i)$ , the conditional expectation can be written as

$$E \left[ t_j e^{-\sum_{i=1}^n t_i^2} | x \right] = E \left[ t_j e^{-t_j^2} | x \right] \prod_{i \neq j} E \left[ e^{-t_i^2} | x \right]$$

$E \left( t_j e^{-t_j^2} | x \right) = 0$  by assumption and hence the estimating equations both have expected value zero.

■

**Corollary 1** *Assume the same conditions as for the theorem. In addition, if given  $x_i$ , the distribution of  $y_i$  is symmetric about  $\mu(\beta, x_i)$ , then the estimating equation  $\frac{d}{d\beta} O_I(y, x_o; \beta)$  is unbiased.*

**PROOF:** Note that  $t_j e^{-t_j^2}$  for  $t_j = y_j - \mu(\beta, x_i)$  is an odd function. Thus, if we assume that the distribution of the  $y_j$ 's is symmetric about  $\mu(\beta, x_i)$ , or equivalently that  $t_j$  is symmetric about zero, then we are integrating an odd function from  $-\infty$  to  $\infty$ , resulting in a value of zero. From this, we see that given a symmetric distribution of the  $t_j$ 's given the covariates, the conditional expectation and hence the original expectation (7) is equal to zero. Thus, the estimating equation that we have chosen is unbiased for distributions of  $y_j$  which are symmetric about its mean. ■

The implication of this theorem and corollary, then, is that our estimates under the appropriate processes will be consistent. On the other hand, we may want to demonstrate consistency of our results for more general distributions of the  $y_i$ 's. We handle this case next, in the context of a more general context.

It seems reasonable that if the estimating equations are asymptotically unbiased, then our estimates will be consistent as well. To approach the issue of consistency from this direction, consider the theorem of Brown (1985) and Ritov (1987).

**Theorem 3** *Suppose that  $W(\theta, P)$  (such that  $W(\theta, P_\theta) = 0$ ) is estimated by  $W_n(\theta, y)$  and that both of these functions are monotone (in  $\theta$ ??).  $W_n(\cdot, \cdot) = 0$  is the estimating equation for  $\theta$ . The unknown parameters are represented by  $\theta$ .*

Further, suppose:

1. There exists some function  $\psi$  such that

$$\int \psi(y, \theta) dP_\theta(y) = 0, \quad \int \|\psi(y, \theta)\|^2 dP_\theta(y) < \infty$$

and

$$W_n(\theta_0, y) = n^{-1} \sum_{i=1}^n \psi(y_i, \theta_0) + O_p(n^{-1/2})$$

2.  $W(\cdot, P)$  is continuously differentiable with a nonsingular matrix of derivatives  $\dot{W}(P)$  at  $\theta$ .

3. For any  $M$ ,

$$\sqrt{n} \{W_n(\theta_0 + t/\sqrt{n}, y) - W_n(\theta_0, y) - W(\theta_0 + t/\sqrt{n}, P_0)\} \xrightarrow{P} 0$$

for  $C = \{|t| \leq M\}$ .

Under the conditions above, there exists a  $\hat{\theta}$  such that  $W_n(\hat{\theta}, y) = o_p(n^{-1/2})$  and any such  $\hat{\theta}$  is  $\sqrt{n}$  consistent. In particular,

$$\sqrt{n}(\hat{\theta} - \theta_0) = n^{-1/2} \sum_{i=1}^n \dot{W}(P_0)^{-1} \psi(Y_i, \theta_0) + o_p(1)$$

and  $\hat{\theta}$  is asymptotically normal.

For proof of this theorem, see ...

We now apply this theorem to our specialized situation. Let  $y_i$  represent the full vector of responses corresponding to the  $i$ 'th individual.

**Theorem 4** Suppose that we have an unbiased estimating equation which can be expressed as  $\frac{1}{n} \sum g(y_i; \theta)$ . Suppose, also, that there exists a function  $G(y_i; \theta)$  such that  $\frac{d}{d\theta} G(y_i; \theta) = g(y_i; \theta)$  and minimizing the function  $\sum G(y_i; \theta)$  with respect to  $\theta$  produces the same estimates as finding the roots of  $\sum g(y_i; \theta)$ .

Create the estimating function

$$\begin{aligned} W_n(\theta, y) &= -\frac{1}{n} \sum g(y_i; \theta) \exp \{G(y_i; \theta)\} \\ &= \frac{d}{d\theta} \exp \left\{ -\frac{1}{n} \sum G(y_i; \theta) \right\}. \end{aligned} \quad (8)$$

and let

$$W(\theta, P) = -E[g(y_i; \theta)] \exp \{-E[G(y_i; \theta)]\}$$

Assume that conditions (2) and (3) (CAN WE REPLACE THE LATTER CONDITION WITH A SIMPLER CONDITION ON  $G$ ????) of the previous theorem hold. In addition, assume (FIX so that CLT holds) the first two moments of  $G(y_i, \theta)$  are finite.

The roots of (8) will produce consistent estimates of  $\theta$ .

**PROOF:** This is a direct application of the previous theorem. In particular, only proof of condition (1) of the previous theorem is needed.

Note, that using a Taylor series expansion of an exponential, we can write

$$\begin{aligned} \exp \left\{ -\frac{1}{n} \sum G(y_i; \theta_0) \right\} &= \exp \{-E_0[G(y_i; \theta_0)]\} \left[ 1 - \left( \frac{1}{n} \sum G(y_i; \theta_0) - E_0[G(y_i; \theta_0)] \right) \right. \\ &\quad \left. + \left( \frac{1}{n} \sum G(y_i; \theta_0) - E_0[G(y_i; \theta_0)] \right)^2 - \dots \right] \\ &= \exp \{-E_0[G(y_i; \theta_0)]\} \left[ 1 + o_p(n^{-1/2}) \right] \end{aligned}$$

We may then rewrite the estimating equation, evaluated at the true parameter value  $\theta_0$ ,

$$\begin{aligned} \frac{d}{d\theta} W_n(\theta_0; y) &= -\frac{1}{n} \sum g(y_i; \theta_0) \exp \left\{ -\frac{1}{n} \sum G(y_i; \theta_0) \right\} \\ &= -\frac{1}{n} \sum g(y_i; \theta_0) \exp \{-E_0[G(y_i; \theta_0)]\} \left[ 1 + o_p(n^{-1/2}) \right] \\ &= -\frac{1}{n} \sum [g(y_i; \theta_0) \exp \{-E_0[G(y_i; \theta_0)]\}] \\ &\quad - \exp \{-E_0[G(y_i; \theta_0)]\} \frac{1}{n} \sum g(y_i; \theta_0) o_p(n^{-1/2}) \\ &= -\frac{1}{n} \sum [g(y_i; \theta_0) \exp \{-E_0[G(y_i; \theta_0)]\}] + o_p(n^{-1/2}) \end{aligned}$$

ACTUALLY, THE ABOVE 1/2 SHOULD BE REPLACED BY 1, I BELIEVE. Thus, let

$$\psi(y_i; \theta_0) = g(y_i; \theta_0) \exp \{-E_0 [G(y_i; \theta_0)]\}$$

Observe that  $E_0 [g(y_i; \theta_0)] = 0$  and its second moment is finite by assumption. Hence, condition (1) is satisfied. The conclusion of Theorem 3 then follows directly. ■

Now, we need ANOTHER THEOREM demonstrating \*INCOMPLETE\* estimating equation produces consistent estimates!!!! (This can be done by showing that the incomplete objective function converges to an objective function which gives consistent estimates upon maximization.

NEXT, show how these theorems apply to regression (allowing us to get consistent estimates even when the distribution of  $y_i$  is not symmetric about  $\mu(\beta, x_i)$ ) and to quasilielihood for any situation.

## 5 P

oisson Regression Example from Lindsey's book on species count on individual Galapagos islands as a function of elevation and area of that particular island. Elevation is a significant predictor. Out of roughly 20 observations, the measurement for elevation is missing for 4/5 cases. (Lindsey used this example to demonstrate use of the over-dispersion parameter, dropping out the observations with missing covariates.)

Further, part of this section will be devoted to demonstrating how to implement the procedure, including a quick way to estimate

$$\int_{y_i}^{\mu(\beta^{(t)}, x_i)} \frac{y_i - t}{V_i(t)} dt$$

## 6 C

conclusions and Future Research

OK, well just thoughts at this point...

1. What properties of the specific form of our objective functions  $\exp\{something\}$  and later taking log's within the EM algorithm are driving our algorithm. Alternatives?
2. Generalized Estimating Equations
3. Nuisance parameters - how do they affect the analysis. e.g. - parameters within the missing data distribution.
4. There are results pertaining to the optimal estimation equation for the full data. What is the optimal estimating equation for the incomplete data? How is it related to that for the full data?