Two Level Proportional Hazards Models

Jerry J. Maples[1],
326 Thomas Building, University Park, PA 16802, maples@stat.psu.edu

Susan A. Murphy[2],
4092 Frieze Building, Ann Arbor, MI 48109-1285
and
William G. Axinn[3]
ISR-4046, 426 Thompson St., Ann Arbor, MI 48106-1248

Summary

We extend the proportional hazards model to a two level model with a random intercept term and random coefficients. The parameters in the multilevel model are estimated by a combination of EM and Newton-Raphson algorithms. Even for samples of 50 groups, this method produces estimators of the fixed effects coefficients that are approximately unbiased and normally distributed. Two different methods, observed information and profile likelihood information, will be used to estimate the standard errors. This work is motivated by the goal of understanding the determinants of contraceptive use among Nepalese women in the Chitwan Valley Family Study (Axinn, Barber, and Ghimire, 1997). We utilize a two-level hazard model to examine how education and access to education for children covary with the initiation of permanent contraceptive use.

Key words: survival analysis; hazard model; frailty model; random coefficient; multilevel; EM algorithm; semi-parametric likelihood; profile likelihood

---

[1]The Methodology Center and Department of Statistics, Pennsylvania State University

[2]Department of Statistics and Institute for Social Research, University of Michigan

[3]Department of Sociology and Institute for Social Research, University of Michigan

1 Introduction

Grouped data occurs in a wide variety of applications, for example, in smoking cessation trials where individuals are in support groups, in educational trials where children are grouped into schools, in community based studies where individuals are grouped into neighborhoods and in genetic studies in which individuals are grouped into families. In many of these same applications the response is the time until an event, such as the time until relapse, the time until marijuana use initiation, time until use of permanent methods of contraception or time until appearance of a disease. Consequently, there has been considerable interest in survival analysis models for grouped or multilevel data (for example, see Clayton, 1978; Oakes, 1982; McGilchrist and Aisbett, 1991; Yashin, Vaupel, and Iachine, 1995; Sastry, 1997; Sinha and Dey, 1997; Vaida and Xu, 2000). Statisticians have recognized that two individuals within the same group will have responses which are more similar than two individuals in two different groups. This increased similarity in responses may be conceptualized as due to a shared group level covariate (Kreft, Leeuw, and Aiken, 1994).

For example, in the Chitwan Valley Family Study (CVFS), neighborhoods of women are sampled in Chitwan Valley, Nepal (Axinn, Barber, and Ghimire, 1997; Axinn and Barber, 2001). A primary goal of this study is to understand the determinants of changing fertility patterns. Of particular interest is the association between a woman's schooling and the timing, relative to the birth of their first child, of initiation of a permanent method of contraception. Some women initiate contraceptive use quickly, averting potential births, while others initiate use late, after bearing many children. Education is expected to increase the opportunity costs of childbearing, motivating more rapid initiation of contraceptive use. The multiple links between the

spread of education and contraceptive behavior, including both neighborhood level effects of schools and individual level effects of schooling, motivate multi-level models of the variation in contraceptive use timing (Axinn, 1993; Axinn and Barber, 2001). Additionally, although research has made great advances in the measurement of all neighborhood information that may correlate with the timing of initiation of permanent contraceptive use (Axinn, Barber, and Ghimire, 1997), it is clear that women in the same neighborhood are likely to have more similar timing patterns than women across neighborhoods.

To allow dependency of response times within a group, shared frailty models were developed (Clayton and Cuzick, 1985; Vaupel, Manton, and Stallard, 1979; Oakes, 1989; Guo and Rodriguez, 1992; Klein, 1992). The shared frailty model is a multilevel extension of the proportional hazards model (Cox, 1972) whereby a frailty (random intercept) term, which varies from group to group, is introduced in the regression model. In effect, different groups experience the event at proportionately different baseline rates. In the shared frailty model, within group comparisons of hazard rates at different levels of an individual level covariate must be the same across groups. In the CVFS, researchers hypothesized that women who have received formal education initiated a permanent method of contraception at a higher rate than women without formal education (Axinn and Barber, 2001). Furthermore, it is plausible that in some neighborhoods there may be little difference in initiation rates between women of different education levels and in other neighborhoods women of different educational levels may vary greatly in their initiation rate. Thus to consider this type of question, we extend the shared frailty model to allow for random coefficients.

This paper will discuss a two level proportional hazards model that incorporates

random variability in the baseline rate and random coefficients for an individual level covariate. Next, we derive the likelihood and elucidate the assumptions behind the likelihood. Parameter estimates will be based on maximum likelihood using a combination of the EM algorithm and Newton-Raphson. Standard errors will be estimated using two different methods, observed information and profile likelihood information. A simulation study will follow to empirically demonstrate properties of the model. The CVFS data will be used to illustrate the applicability of this two level hazard model. Lastly, advantages, disadvantages and areas in need of further research will be discussed.

2 Two level hazard model

2.1 Hazard model

We formulate a two level hazard model for continuous event time data; this model will allow a non-parametric specification of the baseline hazard function. First, the hazard model will be developed and then the likelihood will be given along with a statement of assumptions. Finally, estimation procedures for point estimates and standard errors will be presented.

Let T_{ij} be the event time for the jth individual ($j = 1, ..., n_i$) nested in the ith group ($i = 1, ..., N$) where the event times are measured with enough precision to assume there are no ties. Furthermore, let $Z_i(t)$ denote a vector of group level covariates at time t, $X_{ij}(t)$ denote a p-dimensional vector of individual level covariates at time t and $R_i = (R_{i0}, R_{i1})$ denote $(p+1)$ dimensional vector of unobserved group level random effects. For a study of duration M, on each individual we observe

the minimum of the event time, T_{ij}, and the censoring time, C_{ij}, and an indicator Δ_{ij}, where $\Delta_{ij} = 1$ if $T_{ij} < C_{ij}$ and 0 otherwise. Thus, for group i, we observe $(T_{ij} \wedge C_{ij}, \Delta_{ij}, \{X_{ij}(t), t \in [0, T_{ij} \wedge C_{ij}]\}, j = 1, ..., n_i, \{Z_i(t), t \in [0, M]\})$ and the groups are assumed to be a random sample from the population.

Note that the above Z is assumed to be measured up to time M. In this paper we consider only group level covariates (Z's) that are measured independently of the sampled members of the group. Thus, the group level covariate exists independently from the individuals of the group and although Z's values will not be used after the last person in the group experiences the event or is censored, the group values do exist.

To facilitate the analysis (in Section 4) of the timing of initiation of a permanent method of contraception in the CVFS data, we require that the hazard model incorporates the following multilevel features. First, the hazard model should allow static and time-varying covariates on both the individual and group level. Additionally, the model should accommodate interactions between the individual and group level covariates. Second, the hazard model should allow the effect of individual level covariates to vary between groups. Third, the baseline hazard rates should be permitted to vary between the groups. Finally, if there is no systematic variation in the hazard rates between the groups after accounting for the observed covariates, the model should reduce to Cox's proportional hazard model.

We assume a proportional hazards model for T_i given R_i:

$$\begin{aligned} \lim_{dt \to 0} dt^{-1} P[t \leq T_{ij} < t + dt | X_{ij}(s), Z_i(s), s \leq t, T_{ij} \geq t, R_i] = \\ \lambda_0(t) \exp\{R_{i0} + R_{i1}^T X_{ij}(t) + \beta_x^T X_{ij}(t) + \beta_z^T Z_i(t)\} \\ = h_{ij}(t) \end{aligned} \tag{1}$$

where $\lambda_0(t)$ is the baseline hazard function. Since the above hazard is conditional on the random effects, the β_x coefficients reflect a comparison of responses within the same or identical groups. The variance of R_{i0} measures the heterogeneity of the baseline rate between the groups. The variance of R_{i1} measures the heterogeneity in the coefficient of the individual level covariates X_{ij} between the groups. We assume that the marginal distribution of the random effects is multivariate normal with mean zero and arbitrary covariance matrix Σ. If Σ is 0, then the responses of individuals within a group are independent and this model reduces to the proportional hazards model. We use normally distributed random effects for two reasons. The normal distribution family is closed under linear combinations and matches our view that unexplained variation in hazard rates between groups is due to a large number of unobserved group level covariates.

In addition to the shared frailty model, other multilevel survival models include Hedeker and Gibbons' (1996) development of a multilevel hazard model for interval-time survival data and Barber, et al. (2000) development of a multilevel hazard model for discrete time survival data. Yashin, et al. (1995) developed a correlated frailty model to study the role of genetics versus environmental factors in influencing individual mortality. Sastry (1997) developed a three-level shared frailty model to study the mortality of children in Brazil. Gilks, et al. (1993), Sinha and Dey (1997), Gustafson (1997) and Sargent (1998) have taken a Bayesian approach to modeling grouped survival data.

Vaida and Xu (2000) also consider the proportional hazards model with random effects in the hazard function

$$h_{ij}(t) = \lambda_0(t) \exp(\beta' X^*_{ij} + W'_{ij} R_i), \tag{2}$$

where X^*_{ij} is the group and individual level covariates and W_{ij} is the covariate vector for the random effects, R_i. There are two main differences between the hazard models in (1) and (2). First, in our hazard model (1), $W_{ij} = (1, X_{ij})$, while Vaida and Xu's model allows for a more general random effects structure. Second, our model allows time varying covariates on both the individual and group level covariates. As in our model, the vector of random effects, R_i, in (2) is assumed to follow a multivariate normal distribution with mean zero and unknown covariance matrix. Vaida and Xu use an EM algorithm with MCMC integration in the E-step to estimate the parameters and they use Louis' formula to obtain the observed information matrix (Louis 1982). Our paper provides an alternative procedure for point estimation and computing standard errors and makes the underlying assumptions explicit.

2.2 Likelihood

We follow the likelihood derivation of Nielsen et al. (1992) to provide a partial likelihood for the parameters, $\beta = (\beta_x, \beta_z)$ (the fixed effects), Σ (covariance matrix of random effects), and $\Lambda_0 = \int_0^t \lambda_0(s)ds$ (the cumulative baseline hazard function). The likelihood derivation and assumptions can be made rigorous by using counting process terminology (Aalen, 1976) and the methods developed by Andersen et al. (1993) and Arjas and Haara (1984). Our primary purpose is to provide an intuitive explanation of the assumptions which are made in addition to classical survival analysis assumptions.

We begin by considering one group only. To keep the notation simple, omit the subscript i, denoting group. Also let Y_j denote the observed minimum of the event and censoring time for an individual, $Y_j = T_j \wedge C_j$. Thus, the observations for a group are $(Y_j, \delta_j, \{X_j(t), t \in [0, T_j \wedge C_j]\}, j = 1, ..., n, \{Z(t), t \in [0, \max_j Y_j]\})$.

First, we make a conditional independence assumption. Let $\mathcal{F}_{t-} = \{R,\ T_j I\{T_j < t\},\ I\{T_j < t\},\ \{X_j(s), s \in [0, T_j \wedge t)\}, j = 1, ..., n,\ \{Z(s), s \in [0, \max_j T_j \wedge t)\})\}$ where $I\{A\}$ is one if the event A is true and zero otherwise. This conditional independence assumption may be expressed in terms of conditional probabilities given $\mathcal{F}_{t-}$. We assume that for each j,

$$P\left[t \leq T_j < t + dt \middle| \mathcal{F}_{t-}\right] \approx h_j(t) I\{T_j \geq t\}\ dt \tag{3}$$

for small dt and where $h_j(t) = \lambda_0(t) \exp\{R_0 + R_1^T X_j(t) + \beta_x^T X_j(t) + \beta_z^T Z(t)\}$ as in the previous section. Thus, conditional on R, the hazard of T_j remains the same whether or not we include information on the other subjects in the group up to time t. This assumption is the analog to the conditional independence assumption made in multilevel linear and nonlinear models, that is, given the random effects, responses within a group are assumed independent (for example, see the conditional likelihoods in Rodriguez and Goldman, (1995, eq. 7) or Hedeker and Gibbons (1994, eq. 2)). Note that our parameterization of h_j involves only the jth individual's covariates and not the covariates of other individuals in the group. Thus, if we believe that individual j''s covariate, given by $X_{j'}$, is predictive of the jth individual's response then $X_{j'}$ needs to be included in the jth individual's covariate vector.

Following the likelihood derivation given by Nielsen et al. (1992) for shared frailty model, we assume:

(3) conditional on $R = r$, censoring is independent,

(4) conditional on $R = r$, censoring is noninformative of r and

(5) conditional on $R = r$, the covariates are noninformative of r.

These three assumptions plus the conditional independence assumption (3) and the conditional proportional hazards assumption (1) imply that the group's contribution to a partial likelihood for $(\beta, \Sigma, \Lambda_0)$ is:

$$L(\beta, \Sigma, \Lambda_0) = \int_{\mathcal{R}^{p+1}} \left[\prod_{j=1}^{n} h_j(Y_j)^{\delta_j} \exp\left\{ - \int_0^{Y_j} h_j(s) ds \right\} \right] \times (2\pi)^{-(p+1)/2} |\Sigma|^{-1/2} \exp\{-\frac{1}{2} \mathbf{r}^T \Sigma^{-1} \mathbf{r}\} d\mathbf{r}.$$

To form the partial likelihood for the N groups we subscript $L(\beta, \Sigma, \Lambda_0)$ by i to denote the contribution by group i and multiply across groups to get,

$$L(\beta, \Sigma, \Lambda_0) = \prod_{i=1}^{N} L_i(\beta, \Sigma, \Lambda_0).$$

Assumption (3), independent censoring, is commonly made in the estimation of single level hazard models and has been discussed by many authors (Kalbfleisch and Prentice, 1980; Andersen et al., 1988; Liang, Self, and Chang, 1993). To extend this work to the multilevel setting we make the additional assumptions (4) and (5). Assumptions (4) and (5) are surprisingly stringent. They concern the conditional distribution given the past of the censoring process and the covariate process respectively. To illustrate the stringent nature of these assumptions, we focus on assumption (5). If all of the covariates are time independent then assumption (5) is simply that the covariates are independent of R. Marginal independence between the covariates and the random effects is commonly made in multilevel analyses (Bryk and Raudenbush, 1992; Guo and Rodriguez, 1992; Hedeker and Gibbons, 1996). Define $\mathcal{F}_{t-}^{obs}$ to be $\mathcal{F}_{t-}$ but adjusting for the loss of information due to censoring, that is $\mathcal{F}_{t-}^{obs} = \{R,\ Y_j I\{Y_j < t\},\ \delta_j I\{Y_j < t\},\ I\{Y_j < t\},\ \{X_j(s), s \in [0, Y_j \wedge t)\}, j = 1, ..., n,\ \{Z(s), s \in [0, \max_j Y_j \wedge t)\})\}$. Suppose all time dependent covariates are

exogenous in the sense that the conditional proportional hazards assumption, conditional independence assumption and assumptions (3) and (4) continue to hold even if we include the entire history of the covariate over the interval $[0, M]$ in $\mathcal{F}_{t-}, \mathcal{F}_{t-}^{obs}$. Then as before, assumption (5) is simply that the covariates are marginally independent of R (we have included the entire history of the covariate over the interval $[0, M]$ in $\mathcal{F}_{t-}, \mathcal{F}_{t-}^{obs}$).

Assumption (5) is most stringent when the time dependent covariates do not satisfy the above exogeniety conditions; for example, marginal independence of R and X does not imply that assumption (5) is satisfied. This is because assumption (5) requires conditional independence of R and $X(t)$ given past observations on T. In order to ensure that (5) is a reasonable assumption, we should try to include all common correlates of both the T_j's and the covariates in our model and similarly in order to ensure that (4) is a reasonable assumption, we should try to include all common correlates of both the T_j's and the censoring times. In the appendix we illustrate how marginal independence of R and X is insufficient for assumption (5) to hold.

2.3 Estimation

To estimate the parameters, we maximize an empirical version of the partial likelihood as in Nielsen et al. (1992) and Murphy and van der Vaart (2000). We replace the $\lambda_0(t)$ terms by jumps in the cumulative hazard, $\Delta\Lambda_0(t) = \Lambda_0(t) - \Lambda_0(t-)$. The empirical version of the partial likelihood is then,

$$L(\beta, \Sigma, \Lambda_0) = \prod_{i=1}^{N} \int_{\mathcal{R}^{p+1}} \left[\prod_{j=1}^{n_i} \left(\Delta\Lambda_0(Y_{ij}) e^{g_{ij}(Y_{ij})} \right)^{\delta_{ij}} \exp\left\{ - \int_0^{Y_{ij}} e^{g_{ij}(s)} \, d\Lambda_0(s) ds \right\} \right]$$

$$\times (2\pi)^{-(p+1)/2}|\Sigma|^{-1/2}\exp\{-\frac{1}{2}\mathbf{r}^T\Sigma^{-1}\mathbf{r}\}d\mathbf{r}, \quad (4)$$

where $Y_{ij} = T_{ij} \wedge C_{ij}$, and $g_{ij}(t) = r_{i0} + r_{i1}^T X_{ij}(t) + \beta_x^T X_{ij}(t) + \beta_z^T Z_i(t)$. We will maximize (4) over all non-decreasing Λ_0, all positive definite Σ and real valued vectors β_x and β_z. For any fixed value of (β, Σ) the above partial likelihood will be maximized for $\Lambda_0(t)$, a non-decreasing function in t with positive jumps at the observed event times.

The partial likelihood (4) is maximized over $(\beta, \Sigma, \Lambda_0)$ by using a combination of EM (Dempster, Laird, and Rubin, 1977) and Newton-Raphson. By using both EM and Newton-Raphson methods, the weaknesses of either algorithm are greatly diminished. We use the EM algorithm to first move close to a maximum and then switch to Newton-Raphson to take advantage of its quicker convergence rate. In general, maximizing (3) by EM requires the estimation of the conditional expectations $E[r_i r_i^T | y, \tilde{\theta}]$ and $E[e^{r_{i0}+X_{ij}r_{i1}} | y, \tilde{\theta}]$ where $\tilde{\theta}$ is the current estimates of $(\beta, \Lambda_0, \Sigma)$. We estimate these integrals using a 13 point Gauss-Hermite quadrature. Vaida and Xu (2000) also use the EM algorithm but instead of using a Gauss-Hermite quadrature they use MCMC to approximate the integrals. Once the expectations are formed, the M-step for β follows the usual log partial likelihood in the Cox model with known offsets and $\Delta\Lambda_0(t)$ is estimated by

$$\Delta\Lambda(t) = \left\{ \sum_{y_{ij} \geq t} E[e^{g_{ij}(y_{ij})} | y, \tilde{\theta}] \right\}^{-1}.$$

2.4 Standard Errors

Standard errors can be estimated by at least two different methods, observed information and profile likelihood information. The estimated observed information is the

negative of the Hessian matrix,

$$\hat{I}_{\text{obs}} = -\frac{\partial^2}{\partial\theta\theta^T}\log L(\hat{\theta}),$$

where $\theta = (\beta, \Lambda_0, \Sigma)$. This matrix is used by the Newton-Raphson algorithm for parameter estimation. The asymptotic covariance matrix is then obtained by inverting $\hat{I}_{\text{obs}}$. Since the dimension of the observed information matrix is greater than the number of observed events in the data, it can be quite large.

An alternative method to estimate the standard errors is to use the observed profile likelihood information as in Murphy and van der Vaart (1996). A profile likelihood for β is:

$$\text{Prlik}(\beta) = \max_{\Lambda_0,\Sigma} \log L(\theta).$$

This is treating Λ_0 and Σ as nuisance parameters with respect to inference concerning β. An estimator of the asymptotic covariance matrix of $\hat{\beta}$ is obtained by inverting the observed profile information matrix,

$$\hat{I}_{\text{prof}} = -\frac{\partial^2}{\partial\beta'\beta^T}\text{Prlik}(\hat{\beta}).$$

In a similar fashion we can form a profile likelihood for each component in Σ and then use the observed profile information matrix to form estimators of the asymptotic variance for the estimators of the components in Σ. Patefield (1977) has shown that the standard errors based on the full likelihood and the profile likelihood are equivalent for the reduced parameter space in parametric models. Although the profile likelihood method does not have an explicit expression, we approximate the second derivative through a combined forward and backward finite difference (Murphy and van der Vaart, 1996).

The maximum likelihood estimates and their standard errors (from either the profile likelihood information or the observed information) can be used to create z-statistics. In the next section, we assess how well the distribution of the z-statistics can be approximated by a normal distribution.

3 Simulations

In this section, a simulation study is used to assess the performance of the point estimators and standard errors. Because the primary use of the estimated standard errors is to construct a z-statistic or confidence interval, the accuracy of the standard errors is assessed indirectly by the examining the coverage level of confidence intervals based on the estimated standard errors. All of the simulations use the hazard model $h_{ij}(t) = \lambda_0(t) \exp\{R_{i0} + (\beta + R_{i1})X_{ij}\}$, where X_{ij} is a single individual level static covariate. We consider parameter values that lead not only to nonproportionality but also crossing of the marginal hazard rates.

Each simulation consists of 1000 generated datasets of 50 groups, with each group having uniformly random 4-10 subjects. We consider two covariate distributions, a Bernoulli(.5) and a skewed distribution–$X = E - 1$ where $E \sim$ Exponential(1). The failure time distribution is a Weibull,

$$f(t) = pt^{p-1}e^{r_0+(\beta+r_1)x} \exp\{-t^p e^{r_0+(\beta+r_1)X}\}.$$

We use two values of p; $p = 1$ (flat baseline hazard, $\lambda(t) = 1$) or p=1.5 (increasing baseline hazard, $\lambda(t) = \sqrt{t}$). For each group, the values for the random effects are drawn from a bivariate Normal(0,Σ). The variance-covariance matrix, Σ, will have equal entries on the diagonal. Two censoring levels, 10% and 20% are considered. The

event times are censored at the 90^{th} or 80^{th} percentile of the event times distribution conditional on the covariate.

Table 1 gives the parameters values for all six representative simulations. For each of the 1000 simulated datasets, 95% confidence intervals were created for each parameter using the profile likelihood information standard errors. The mean, empirical 95% confidence interval and coverage probability are listed in Table 2.

[Insert Table 1. about here]

Simulations 1-4 represent a variety of situations in which we expect the point estimates and estimated standard errors to perform well. Since each simulation consists of 1000 datasets, we expect the coverage percentages to be between 93.6% and 96.4%. Only simulation 4 yields a coverage percent for β that falls outside of this interval. In this case increasing the sample size to 100 yields a coverage percent falling in the interval. Since all of the 50-group simulations had a coverage slightly less than 95% and the point estimators appear to be unbiased, this may be an indication that in small samples the profile likelihood information slightly underestimates the standard error for β. In all four simulations, the standard deviations of the random effects are underestimated. The downward bias was typically larger for the random coefficient than the random intercept. This bias can be expected because the estimators do not take into account the loss in degrees of freedom from the estimation of β (Fahrmeir and Tutz, 1994, chapter 7).

[Insert Table 2. about here]

For simulations 5 and 6, the assumptions behind the model are violated and thus examine the robustness of the model. In simulation 5, $\Sigma = \mathbf{0}$, the variance matrix

for the random effects is on the boundary of the parameter space, and thus we do not expect the point estimates and estimated standard errors to behave well. In simulation 6, the setup is again identical to simulation 1 except that the random effects are generated from independent demeaned gamma distribution $W - 1$, where W has mean 1 and variance .1827. Simulation 5 showed the most problems, which was to be expected. However, the estimator of β had only a slight positive bias and the approximate 95% confidence intervals performed well. Unlike the negative bias exhibited by the estimated standard deviations in simulations 1-4 and 6, the estimated standard deviations are positively biased in simulation 5. This positive bias is expected because the estimation algorithm forces the parameter estimates to stay within the interior of the parameter space. With the exception of the covariance term, the point estimates and estimated standard errors performed well in simulation 6, even when the random effects distribution was misspecified.

In Table 3, a comparison between the two standard error estimation methods, profile likelihood information and observed information, is made. We focus on the estimated standard errors for β and compare the coverage of the 95% normal theory confidence intervals. The two methods have nearly identical coverage properties, however, in some of the datasets where the parameter estimates for the covariance matrix of the random effects were on the boundary of the parameter space, the Hessian matrix was not positive definite.

[Insert Table 3. about here]

4 Chitwan Valley Family Study

The Chitwan Valley Family Study (CVFS) was designed to measure dynamic changes in socioeconomic context from a sample of 171 neighborhoods in South-Central Nepal, and to link these changes to individual level life histories for the purpose of explaining marriage timing, childbearing, and contraceptive use (Axinn, et al., 1997). Neighborhoods are defined as clusters of five to fifteen households, which fits the settlement pattern in Chitwan (Axinn, et al., 1997). The retrospective histories of change in each neighborhood were collected with the aid of the Neighborhood History Calendar method (Axinn, et al., 1997). In each sampled neighborhood, CVFS interviewed every resident aged 15 to 59. Residents were asked to provide complete life histories of childbearing, contraceptive use, education and related behaviors. Building on on the Life History Calendar method (Freedman et al., 1988), the investigators developed an advanced form of the life history calendar, including memory cues from the neighborhood history data (Axinn, Pearce, and Ghimire, 1999).

A central aim of the CVFS was the examination of the timing of initiation of permanent contraceptive use. The vast majority of contraception in Nepal is used for stopping childbearing rather than spacing births. For the purposes of this example, we consider initiation of the use of IUD, Norplant, and depo-provera to be permanent methods (see Axinn and Barber, 2001 for a more complete list and discussion). Because permanent contraceptive use among women who have no children is extremely rare in this setting, we will only estimate rates of initiation of permanent contraceptive use for women who have given birth to at least one child. The time of first use of a permanent method of contraception is measured in months from the birth of a woman's first child. If a woman had not used contraceptives at the time of

the interview, then the censoring time was measured in months from the birth of the woman's first child to the interview.

We use this study to illustrate the usefulness of the proposed methodology. More complete multilevel analyses can be found in Axinn and Barber (2001) and Axinn and Yabiku (2001). Because this particular substantive application is designed as an illustration of the multilevel model, we do not provide a comprehensive examination of other potential methodological issues relevant to the substantive subject, such as problems of retrospective recall (see Diamond, McDonald, and Shah, 1986) or bias due to differential death rates.

Access to schooling for children, the level of education of the woman and birth cohort are among the factors that may predict the rate at which women start to use contraceptives. The main hypotheses that will be tested here are:

hypothesis 1 Do women with some formal education have a higher rate of initiating a permanent method of contraception compared to uneducated women within their neighborhood?

hypothesis 2 Does the effect of women's education on the time of initiation of a permanent method of contraception vary across the neighborhoods?

hypothesis 3 Do women who are located in neighborhoods closer to schools have a higher rate of initiating a permanent method of contraception compared to women in neighborhoods further away from schools?

We considered a sample of 81 neighborhoods, all within 9.7 miles of the Narayanghat, the major town in Chitwan valley. Within the neighborhoods, we considered all women between the ages of 25-44 (born between 1952-1971) at the time of the 1996 in-

terview, each of whom had given birth to at least one child. This sample is composed of 488 ever-married women with a range of 1 to 15 women per neighborhood.

In the simple model below we use three covariates. The indicator [COHORT] is equal to one for the older women, who were between the ages of 35-44 during the interview (born between 1952-1961), and equal to zero for the younger women (born between 1962-1971). If the woman had any formal education prior to the birth of her first child then the indicator for education, [EDU], equaled one. In the older cohort of women, 55 out of 196 (28.1%) had received any formal education prior to the birth of their first child. For the younger cohort of women, 160 out of 292 (54.8%) had received any formal education. To measure access to educational opportunities for children, the indicator [DIST-SCH] was equal to one if if there was a school within a 5 minute walk of the neighborhood and zero otherwise. The distance to nearest school is a time-varying covariate at the neighborhood level. Prior to the 1950s, public schools were extremely rare in the Chitwan valley area, but since, there has been a dramatic increase in the number of schools, thus increasing the educational opportunities for children.

Assumptions (3) and (4) from Section 2.2 are trivially satisfied as censoring occurs at the number of months corresponding to the length of time from the birth of the woman's first child to the time of the interview. Consider assumption (5). The only time dependent covariate, distance to nearest school, is a neighborhood level variable. Since decisions concerning school placement were made by the central government, we believe that this covariate is exogenous in the sense that the distance from a neighborhood to a school is independent of past actions by women in the neighborhood. Thus conceptually we may include the entire history of the distance-

to-nearest school covariate in $\mathcal{F}_{t-}, \mathcal{F}_t^{obs}$, i.e., in terms of forming the likelihood we treat distance to nearest school as a time independent covariate. So assumption (5) reduces to assuming that the random effect R is marginally independent of all of the covariates.

In order to address the three hypotheses, we fit the following hazard model:

$$h_{ij}(t) = \lambda_0(t) \exp\{R_{0i} + (\beta_1 + R_{1i})\text{EDU}_{ij} + \beta_2 \text{COHORT}_i + \beta_3 \text{EDU-COHORT}_{ij} + \beta_4 \text{DIST-SCH}_i(t)\}$$

The parameter estimates and standard errors are presented in Table 4. All of the covariates were significant in the model. From the model and Table 4, we can address the three main hypotheses for this analysis. Since the interaction between birth cohort and the woman's education was significant, we must address hypothesis 1 for each cohort separately. In the younger cohort, the women with some formal education prior to the birth of their first child had a hazard rate $e^{\hat{\beta}_1} = e^{.579} = 1.78$ times higher than women without any education (p-value <.0001). However, for the older cohort of women, the hazard rate was $e^{\hat{\beta}_1 + \hat{\beta}_3} = e^{.091} = 1.09$ (S.E.$(\hat{\beta}_1 + \hat{\beta}_3)$=.059, p-value = .1231) times higher than the older women without any formal education, but this was not a significant difference. For a more complete discussion, including consideration of the availability of contraceptive methods, and the relation between education and the use of permanent contraceptives see Axinn and Barber (2001).

The variance for the random coefficient (R_{1i}), for women's education, used to test the second hypothesis, is nonsignificant (p-value .2005), providing evidence that the covariation of education with initiation of a method of permanent contraception is constant across neighborhoods. From the third hypothesis, women who lived in neighborhoods with schools nearby, within a 5 minute walk, had a 21% higher hazard

rate ($e^{\hat{\beta}_4} = e^{.193} = 1.21$) for initiation than women in neighborhoods that did not have a school nearby(p-value $<.0001$). In addition to the main hypotheses, the model also gives further insight into the data. The variance of the baseline hazard rate between the neighborhoods, $\mathrm{Var}(R_{0i})$, is significant (p-value $<.0001$). We will discuss credibility of these inferences, in particular the credibility of the estimated standard errors, using profile likelihood surface plots.

[Insert Table 4. about here]

Examining the profile likelihood surface for a pair of parameters can show how well the normal theory Z-tests are working. If the normality approximation holds, then the contours of the profile likelihood surface will be elliptical. The profile likelihood surface plot for (β_1, β_2), the effects of woman's education and cohort, is shown in Figure 1. The contours represent twice the difference in profile likelihood value from the maximum likelihood. The elliptical shape of the contour lines indicate that the standard errors should work well, i.e. the 95% confidence set for (β_1, β_2) is quadratic. The profile likelihood surface plot for (std.dev(R_0),std.dev(R_1)) is shown in Figure 2. In this plot, the contour lines do not have the nice elliptical shapes as before. Viewing the plot from the y axis, it is clear that the quadratic approximation of the likelihood does not hold for estimation of the standard deviation of the random coefficient; however viewing the plot from the x axis, we see that the surface of the profiled likelihood does appear quadratic in the standard error of the random intercept. Certainly some skepticism is in order in interpreting the p-values for the variance components.

[Insert Figures 1. and 2. about here]

5 Discussion

The two level proportional hazards model allows for heterogeneity between groups as well as for effects of individual level covariates to vary by group, i.e. random coefficients. The fixed effects represent comparisons of individuals within the same or highly similar groups, rather than a comparison between individuals from a variety of neighborhoods. In contrast to the frailty model, this model allows us to examine the level of covariation between a subject level covariate such as woman's education and neighborhood. A combination of EM and Newton-Raphson algorithms are used to obtain parameter estimates. Standard errors may be estimated using the observed information or a finite difference of the profile likelihood.

When the fixed effect, β, is of primary scientific interest, the simulations in Section 3 provide evidence for the usefulness of this model and estimation method, for datasets as small as 50 groups. The estimators and standard errors for the variance components need improvement, especially the estimator of the covariance term. It is easy to see how the model can be extended to accommodate any number of random coefficients and more complicated random effects design. However, computational issues in point estimation and standard errors must be carefully considered. The development of methods other than the traditional EM and Newton-Raphson are needed for these high dimensional maximization problems. Alternatives include both Vaida and Xu's (2000) use of MCMC integration and Raudenbush, Yang, and Yosef's (2000) use of a multivariate Laplace approximation to the likelihood function. Further, asymptotic theory for the estimators in this model still needs to be developed. Finally, formal testing of the variance of the random effects being equal to zero needs to be explored, as this is often an important research hypothesis.

Acknowledgments. This research was supported by grant 1P50-DA-10075 from NIDA to the Pennsylvania State University's Methodology Center, grant HD32912 from NICHD, NSF grants SBR 9811983 and DMS 9802885 and a P30 center grant from NICHD to the Institute for Social Research at University of Michigan.

Appendix: Example for Assumption 5

The following example illustrates how R and X may be marginally independent yet given observation on T, R and X are dependent. There is no censoring. The covariate, $X(t)$ is identically zero (with probability one) for $t \leq t_0$ and is discrete valued thereafter. Furthermore, suppose that there is an unobserved binary common correlate, say U, of both T and $X(t_0 + dt)$. The joint density of $(R, U, S = I\{T \geq t_0\}, X(t_0 + dt))$ evaluated at (r, u, s, x) can be written as,

$$f_R(r)P[U = u]P[T \geq t_0|R = r, U = u]^s P[T < t_0|R = r, U = u]^{1-s} \times P[X(t_0 + dt) = x|U = u],$$

where f_R is the density of R. To be concrete, let $P[U = 0] = P[U = 1] = 1/2$, $P[T \geq t_0|R = r, U = 1] = 2e^{-e^r t_0} - e^{-2e^r t_0}$ and $P[T \geq t_0|R = r, U = 0] = e^{-2e^r t_0}$. It is easy to see that X is marginally independent of R and U is marginally independent of the random effect, R. The hazard model constrains the form of $P[T \geq t_0|R = r] = \sum_u P[U = u]P[T \geq t_0|R = r, U = u]$ (in this case $P[T \geq t_0|R = r] = e^{-e^r t_0}$) and specifies that f_R is a normal distribution; the other assumptions impose no further constraints. Because both U and R are causes of T, we can expect that U and R are correlated given $T \geq t_0$. That is,

$$P[U = 0|T \geq t_0,\ R = r] = \frac{1/2e^{-2e^r t_0}}{e^{-e^r t_0}}$$

depends on $R = r$. However U is also correlated with $X(t_0 + dt)$. Thus we can expect

that $X(t_0 + dt)$ and R are correlated given $T \geq t_0$. That is,

$$P[X(t_0+dt) = x | T \geq t_0,\ R = r] = \frac{1/2 e^{-2e^r t_0} P[X(t_0+dt) = x | U = 0]}{e^{-e^r t_0}} + \frac{1/2(2e^{-e^r t_0} - e^{-2e^r t_0}) P[X(t_0+dt) = x | U = 1]}{e^{-e^r t_0}}$$

will generally depend on $R = r$. In this case assumption (5) will be violated.

References

Aalen, O. (1976) Nonparametric Inference in Connection with Multiple Decrement Models. *Scandinavian Journal of Statistics* **3**, 15-27.

Andersen, P., Borgan, O., Gill, R. and Keiding, N. (1988). Censoring, Truncation and Filtering in Statistical Methods Based on Counting Processes. *Contemporary Mathematics* **80**, 19-60.

Andersen, P., Borgan, O., Gill, R. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag.

Arjas, E. and Haara, P. (1984). A marked point process approach to censored failure data with complicated covariates. *Scandinavian Journal of Statistics* **11**, 193-209.

Axinn, W. (1993). The Effects of Children's Schooling on Fertility Limitation. *Population Studies* **47**,481-493.

Axinn, W. and Barber, J. (2001) Mass Education and Fertility Limitation. *to appear in American Sociological Review.*

Axinn, W., Barber, J. and Ghimire, D. (1997). *Sociological Methodology.* chapter: The Neighborhood History Calendar, Adrian Raftery, editor. Blackwell Publishers, 355-392.

Axinn, W., Pearce, L. and Ghimire, D. (1999). Innovations in Life History Calendar Applications. *Social Science Research* **28**, 243-264.

Axinn, W. and Yabiku, S. (2001) Social Change, the Social Organization of Families, and Fertility Limitation. *to appear in American Journal of Sociology.*

Barber, J., Murphy, S., Axinn, W. and Maples, J. (2000). Discrete Time Multilevel Hazards Analysis. *Sociological Methodology* **30**, 201-235.

Bryk, A. and Raudenbush, S. (1992). *Hierarchical Linear Models.* Sage Publications.

Clayton, D. (1978). A Model for Association in Bivariate Life Tables. *Biometrika* **65**, 141-151.

Clayton, D. and Cuzick, J. (1985). Multivariate Generalizations of the Proportional Hazards Model. *Journal of the Royal Statistical Society, series A* **148**, 82-117.

Cox, D. (1972). Regression Models and Life Tables. *Journal of the Royal Statistical Society, series B* **34**, 187-220.

Dempster, A., Laird, N. and Rubin, D. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, series B* **39**, 1-38.

Diamond, I.D., McDonald, J.W. and Shah, I.H. (1986). Proportional hazards models for current status data: application to the study of differentials in age at weaning in Pakistan. *Demography* **23**, 607-620.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer-Verlag.

Freedman, D., Thornton, A., Camburn, D., Alwin, D. and Yong-DeMarco, L. (1988). The Life History Calendar: A Technique for Collecting Retrospective Data. In **Sociological Methodology 1988**, edited by Clifford C. Clogg, 37-38.

Gilks, W., Clayton, D., Spiegelhalter, D., Best, N. and McNeil, A. (1993). Modeling Complexity: Applications of Gibbs Sampling in Medicine. *Journal of the Royal Statistical Society, series B* **55**, 39-52.

Guo, G. and Rodriguez, G. (1992). Estimating a Multivariate Proportional Hazards Model for Clustered Data using the EM Algorithm. *Journal of the American Statistical Association* **87**, 969-976.

Gustafson, P. (1997) Large Hierarchical Bayesian Analysis of Multivariate Survival Data, *Biometrics* **53**, 230-242.

Hedeker, D. and Gibbons, R. (1994). A Random-Effects Ordinal Regression Model for Multilevel Analysis. *Biometrics* **50**, 993-944.

Hedeker, D. and Gibbons, R. (1996). MIXOR: a Computer Program for Mixed Effects Ordinal Regression Analysis. *Computer Methods and Programs in Biomedicine* **49**, 157-176.

Kalbfleisch, J. and Prentice, R. (1980). *The Statistical Analysis of Failure Time Data.* Wiley.

Keiding, N., Andersen, P. and Klein, J. (1997). *Statistics in Medicine* **16**, 215-224.

Klein, J. (1992). Semiparametric Estimation of Random Effects Using the Cox Model Based on the EM Algorithm. *Biometrics* **48**, 795-806.

Kreft, I., Leeuw, J. and Aiken, L. (1994). The Effect of Different Forms of Centering in Hierarchical Linear Models. Technical Report 30. National Institute of Statistical Sciences.

Liang, K., Self, S. and Chang, Y. (1993). Modeling Marginal Hazards in Multivariate Failure Time Data. *Journal of the Royal Statistical Society, Series B* **55**, 441-453.

Louis, T. (1982). Finding Observed Information using the EM Algorithm. *Journal of the Royal Statistical Society, Series B* **44**, 98-130.

McGilchrist, C and Aisbett, C. (1991). Regression with Frailty in Survival Analysis. *Biometrics* **47**, 461-466.

Murphy, S. and van der Vaart, A. (1996). Semiparametric Likelihood Ratio Inference, Technical Report 96-03. Pennsylvania State University. Department of Statistics.

Murphy, S., van der Vaart, A. (2000). On Profile Likelihood. *Journal of the American Statistical Association* **95**, 449-485.

Nielsen, G., Gill, R., Andersen, P. and Sorensen, T. (1992). A Counting Process Approach to Maximum Likelihood Estimation in Frailty Models. *Scandinavian Journal of Statistics* **19**, 25-43.

Oakes, D. (1982). A Model for Bivariate Survival Data. *Journal of the Royal Statistical Society, series B* **44**, 414-422.

Oakes, D. (1989). Bivariate Survival Models Induced by Frailties, *Journal of the American Statistical Association* **84**, 487-493.

Patefield, W. (1977). On the Maximized Likelihood Function. *Sankhya, series B* **39**, 92-96.

Raudenbush, S., Yang, M. and Yosef, M. (2000). Maximum Likelihood for Generalized Linear Models with Nested Random Effects via High-Order, Multivariate Laplace Approximation. *Journal of Computational and Graphical Statistics* **9**, 141-157.

Rodriguez, G. and Goldman, N. (1995). An Assessment of Estimation Procedures for Multilevel Models with Binary Responses. *Journal of the Royal Statistical Society, Series A* **158**, 73-89.

Sargent, D. (1998). A General Framework for Random Effects Survival Analysis in the Cox Proportional Hazards Setting. *Biometrics* **54**, 1486-1497.

Sastry, N. (1997). A Nested Frailty Model for Survival Data. *Journal of the American Statistical Association* **92**, 426-435.

Sinha, D. and Dey, D. (1997). Semiparametric Bayesian Analysis of Survival Data. *Journal of the American Statistical Association* **92**, 1195-1212.

Vaida, F. and Xu, R. (2000). Proportional Hazards Model with Random Effects. *Statistics in Medicine* **19**, 3309-3324.

Vaupel, J., Manton, K. and Stallard, E. (1979). The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality. *Demography* **16**, 439-454.

Yashin, A., Vaupel, J. and Iachine, I. (1995). Correlated Individual Frailty: An Advantageous Approach to Survival Analysis of Bivariate Data. *Demography* **34**, 31-48.

Table 1: Simulated datasets

Sim. No.	β	Censor Rate	Hazard Shape	Var. R.E.	Corr.	Covariate
1	1	10%	flat	.1827	0	Bernoulli
2	1	10%	flat	1	0	Bernoulli
3	.25	20%	increasing	.4112	.5	Bernoulli
4	-.25	20%	increasing	.4112	-.5	Exponential
5	1	10%	flat	0	0	Bernoulli
6	1	10%	flat	.1827	0	Bernoulli

Parameter values for the simulation study. In simulation 6, the random effects come from a demeaned Gamma distribution rather than a normal distribution. Simulation 2 has converging hazard functions. Simulations 3 and 4 have crossing hazard functions.

Table 2: Simulation Results

Sim. No.	No. Groups	Parameter	True Value	Mean Value	95% Conf. Int.	Coverage Percent
1	50	β	1	1.007	(.997,1.016)	.946
		s.d.(int)	.427	.409	(.401.418)	.947
		cov	0	-.008	(-.015,-.002)	.986
		s.d.(coef)	.427	.395	(.382,.407)	.943
2	50	β	1	.999	(.987,1.012)	.939
		s.d.(int)	1	.970	(.959,.980)	.937
		cov	0	.023	(.007,.039)	.938
		s.d.(coef)	1	.947	(.934,.961)	.954
3	50	β	.25	.253	(.242,.263)	.941
		s.d.(int)	.641	.629	(.621,.638)	.961
		cov	.206	.185	(.186,.204)	.982
		s.d.(coef)	.641	.615	(.602,.629)	.919
4	50	β	-.25	-.242	(-.251,-.235)	.925
		s.d.(int)	.641	.620	(.613,.627)	.980
		cov	-.205	-.202	(-.209,-.195)	.966
		s.d.(coef)	.641	.611	(.602,.620)	.919
4(a)	100	β	-.25	-.249	(-.256,-.242)	.953
		s.d.(int)	.641	.627	(.620,.633)	.925
		cov	-.205	-.199	(-.207,-.191)	.923
		s.d.(coef)	.641	.618	(.608,.627)	.983
5	50	β	1	1.017	(1.008,1.026)	.945
		s.d.(int)	0	.155	(.147,.162)	.913
		cov	0	-.004	(-.005,-.003)	.992
		s.d.(coef)	0	.231	(.221,.242)	.869
6	50	β	1	.996	(.987,1.006)	.948
		s.d.(int)	.427	.410	(.401,.419)	.938
		cov	0	-.018	(-.025,-.011)	.988
		s.d.(coef)	.427	.400	(.387,.413)	.935

Table 3: Comparison of Standard Errors for β

Sim. No.	True Value	Coverage (Prof. Like.)	Coverage (Obs. Info.)	Number of Datasets Obs. Info. Problems*
1	1	.946	.946	7
2	1	.939	.938	6
3	.25	.941	.942	17
4	-.25	.925	.920	2
4(a)	-.25	.953	.953	3
5	1	.945	.947	29
6	1	.948	.946	10

*Datasets in which the Hessian matrix used in calculating the observed information was not positive definite.

Table 4: Results from Analysis of CVFS

	Estimate	Std. Error	Z-score (P-value)
Covariates			
Education	.579	.0461	12.56 (<.0001)
Cohort	-.121	.0473	2.55 (.0108)
Edu-Cohort	-.488	.0456	10.70 (<.0001)
Dist-School	.193	.0372	5.18 (< .0001)
Variance Components			
Std. Dev (r_{0i})	.571	.128	4.46 (<.0001)
Std. Dev (r_{1i})	.381	.297	1.44 (.2005)
Covariance	-.132	.169	0.781 (.4348)

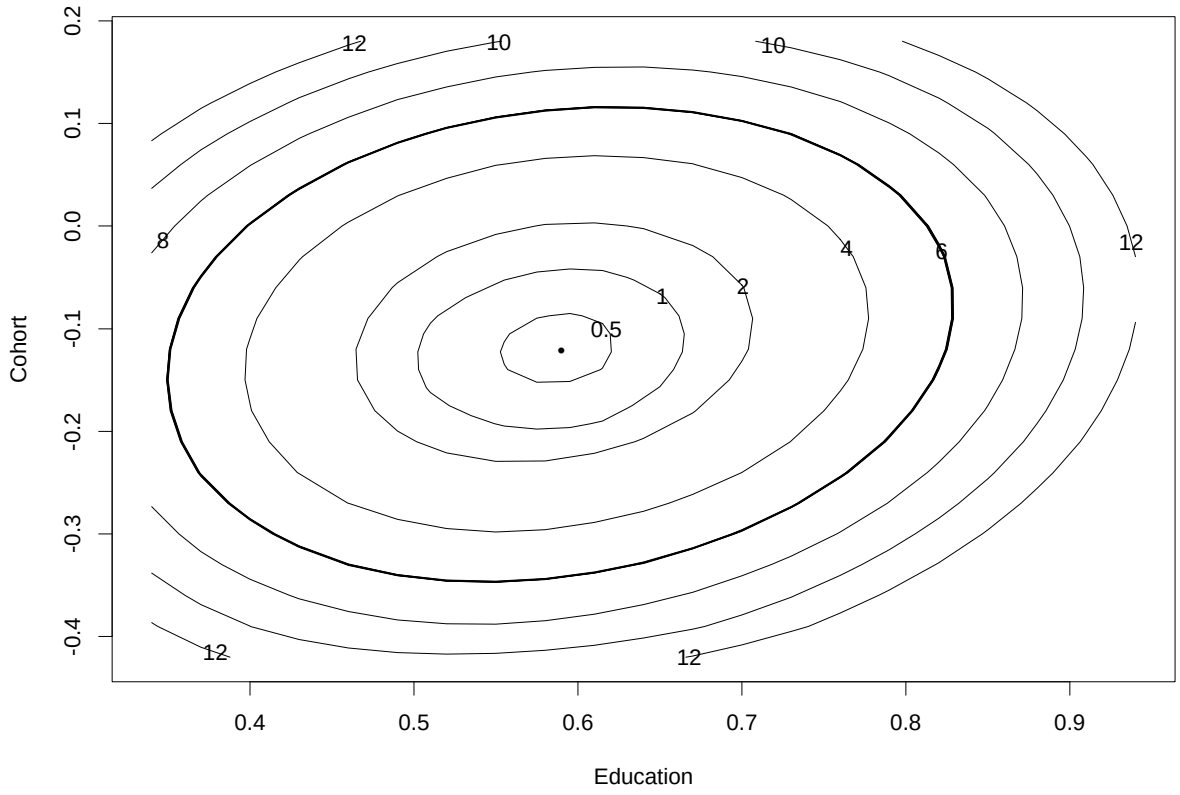

Figure 1: Profile likelihood surface for parameters β_1 (Education) and β_2 (Cohort). The contour lines represent 2 times the difference in likelihood value from the maximum likelihood value at the MLE. Assuming normality, the darkened contour line of 6 does represents a 95% joint confidence region for β_1 and β_2.

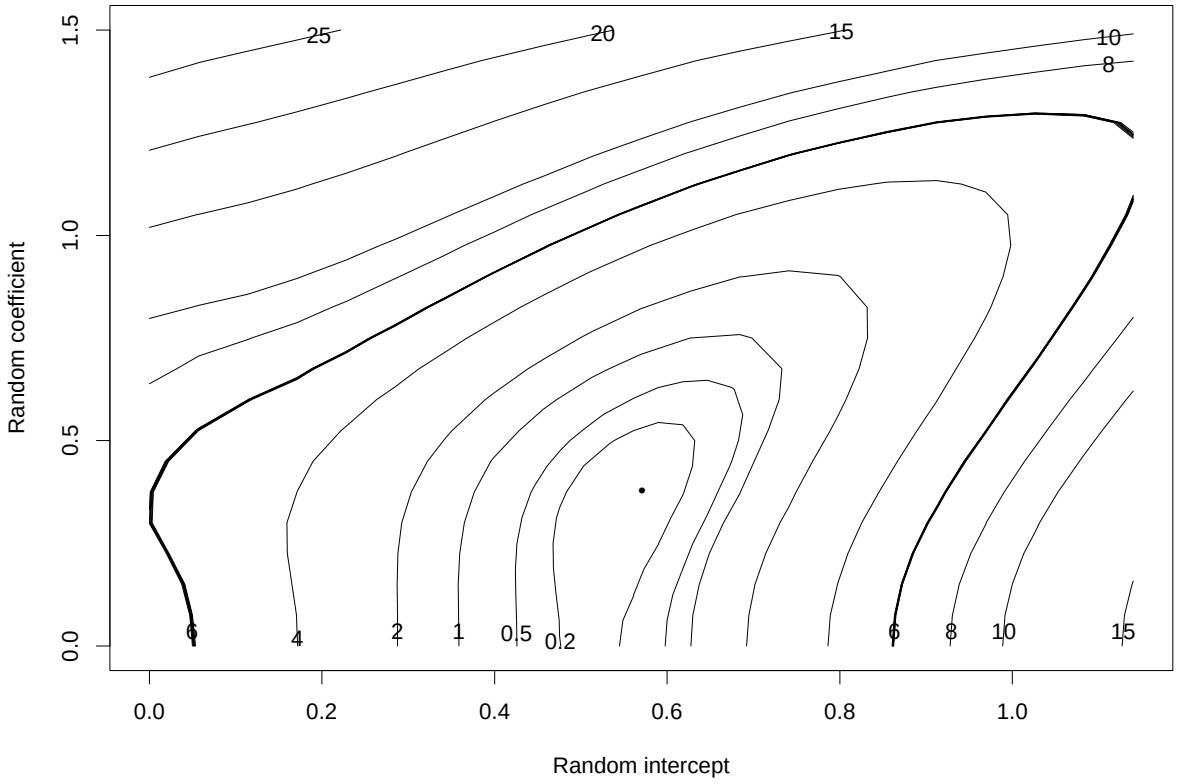

Figure 2: Profile likelihood surface for standard deviations of r_0 (random intercept) and r_1 (random coefficient). The contour lines represent 2 times the difference in likelihood value from maximum likelihood value at the MLE. Assuming normality, the darkened contour line of 6 represents a 95% joint confidence region for the standard deviations of r_0 and r_1.