Mixed Effects Cox Models

Terry Therneau
Mayo Clinic

June 15, 2015

1 Introduction

Like many other projects, this software library started with a data set and a problem. From this came statistical ideas for a solution, followed by some initial programming — which more than anything else helped to define the *real* computational and statistical issues — and then a more ambitious programming solution. The problem turned out to be harder than I thought; the first release-worthy code took over 3 years in gestation and resulted in the *kinship* library. This in turn led to application to a larger set of problems, and a complete re-design of the underlying code. It was then split into separate libraries: coxme contains the central code for mixed effects Cox and linear models, kinship2 contains code for the construction and manipulation of pedigree data, and bdsmatrix contains underlying routines for block-diagonal sparse matrices.

2 Model

The coxme function fits the model

$$\lambda(t) = \lambda_0(t)e^{X\beta+Zb}$$
$$b \sim G(0, \Sigma(\theta))$$

where λ_0 is an unspecified baseline hazard function, X and Z are the design matrices for the fixed and random effects, respectively, β is the vector of fixed-effects coefficients and b is the vector of random effects coefficients. The random effects distribution G is modeled as Gaussian with mean zero and a variance matrix Σ, which in turn depends a vector of parameters θ.

For any fixed values of β and b we define the usual Cox partial likelihood PL (often labeled as a log-likelihood in printouts) as

$$\log[PL(\beta,b))] = \sum_{i=1}^{n} \int_0^\infty \left[Y_i(t)\eta_i(t) - \log\left(\sum_j Y_j(t)e^{\eta_j(t)} \right) \right]$$

where $\eta_i(t) = X_i(t)\beta + Z_i(t)b$ is the linear score for subject i at time t and $Y_i(t)$ describes the risk set, $Y_i(t) = 1$ if subject i is still under observation at time t and 0 otherwise. For further details see chapter 3 of Therneau and Grambsch [5] or any other mid-level survival text.

We can integrate out the random effects to create the integrated partial likelihood

$$IPL(\beta, \theta) = \frac{1}{(2\pi)^{q/2}|\Sigma(\theta)|^{1/2}} \int PPL(\beta, b)e^{-b'\Sigma^{-1}(\theta)b/2} \, db$$

where q is the length of b, i.e., the number of random effects. Ripatti and Palmgren [3] showed that the IPL can be treated as a likelihood, just as integrated full likelihoods can. An ML estimate is obtained by joint maximization over β and θ. We can compare a nested set of random or fixed effects Cox models using chi-square tests.

The REML estimate is obtained by also integrating out the fixed effects

$$REML(\theta) = \frac{1}{(2\pi)^{q/2}|\Sigma(\theta)|^{1/2}} \int PPL(\beta, b)e^{-b'\Sigma^{-1}(\theta)b/2} \, db \, d\beta$$

The REML estimate is obtained by maximization of this quantity over θ. As in linear models, the REML estimate cannot be used for testing fixed effects, due to the lack of a proper prior distribution for β. For instance, assume that `age` were a candidate variable. The REML value for a fit with age in days versus one with age in years will differ by the constant $\log(365.25)$; this has no impact on the estimate of θ, but makes the comparison of two models, one with and one without age, completely arbitrary.

In linear models the REML estimate is often preferred, since practical experience has shown it to be more reliable; the ML estimates are often too large. By analogy, some have recommended using a REML estimate for the mixed effects Cox model. However, either theoretical and practical evidence for its superiority is sparse. The coxme function currently only provides an ML estimate.

For a very simple fit such as

```
> fit1 <- coxme(Surv(endage, cancer) ~ parity + (1| famid))
```

found below there will be one random intercept per group (family), and Z will be the usual design matrix for a one-way anova, i.e., the same design matrix as would be used by a linear model `lm(y ~ factor(famid)-1)`. In this case $Z_{ij} = 1$ iff subject i is a member of family j. The variance of the random effects in this case is $\Sigma = \theta I$.

Note that if there are k groups the vector of random effects coefficients b will have k elements. Because of the shrinkage provided by the random effect they will satisfy $\sum b_k = 0$, rather than having one of their members set to zero. With respect to the random effects the `contrasts` option has no relevance.

3   Simple Models

For many random effects models the random effects can be specified very simply in the model formula. As an example consider the `eortc` data set, which is included with the package for illustration. This is a simple simulated example, based on the results of a breast cancer trial undertaken by the European Organization for Research and Treatment of Cancer. There are 37 enrolling centers with enrollments ranging from 21 to 247 subjects. We start by fitting a simple model with a random intercept per center.

```
> library(coxme)
> stem(table(eortc$center))

  The decimal point is 1 digit(s) to the right of the |

   2 | 1335600244455579
   4 | 12232236
   6 | 1362
   8 | 561
  10 | 46
  12 | 0
  14 | 5
  16 |
  18 | 3
  20 |
  22 |
  24 | 7

> efit1 <- coxph(Surv(y, uncens) ~ trt, eortc)
> efit2 <- coxme(Surv(y, uncens) ~ trt + (1|center), eortc)
```

Random effects are specified in the formula by a parenthesised expression which contains a vertical bar separating effects on the left from grouping variables on the right. In the case above we read it as an intercept (effect) per center (group).

```
> print(efit2)

Cox mixed-effects model fit by maximum likelihood
  Data: eortc
  events, n = 1463, 2323
  Iterations= 9 49
                    NULL Integrated     Fitted
Log-likelihood -10638.71  -10520.65 -10478.84

                   Chisq    df p   AIC    BIC
Integrated loglik 236.11  2.00 0 232.11 221.53
 Penalized loglik 319.74 28.69 0 262.37 110.67

Model:  Surv(y, uncens) ~ trt + (1 | center)
Fixed coefficients
         coef exp(coef)   se(coef)      z p
trt 0.7086127  2.031171 0.06424398 11.03 0

Random effects
 Group  Variable  Std Dev   Variance
 center Intercept 0.3292140 0.1083818
```

The components of the printout are

- The total number of observations and the total number of events (deaths) in the data set.

- The computational effort, summarized as the number of iterations for the optim routine and the underlying Newton-Raphson iterations used.

- The log partial likelihood for a model with no covariates or random effects, the fitted partial likelihood, and the value with the random effects integrated out. We will normally be interested in the null and integrated values. (The log values are printed, but labeled as PL and IPL for brevity).

- Likelihood ratio tests based on the integrated and penalized views of the model, along with penalized values. The AIC penalizes by twice the effective degrees of freedom, and the BIC by $\log(d)$ times the effective degrees of freedom, where d is the number of events.

- A summary of the fixed effects

- A summary of the variances of the random effects

After integrating out the random effects, the log partial likelihood for the mixed effects model is 236.1. As a test of the random effects, we would normally compare this to `efit1`, the fit with no random effects which has a log partial likelihood of 105.7 The estimated standard deviation between centers of .33 is fairly substantial (more on this below). The difference is > 100 on one degree of freedom, which is highly significant.

Another way to approach the fit is as a penalized model. The coefficients are viewed as the solution to a penalized problem $\log(PPL(\beta, b)) - \log(PL(\beta, b)) - p(\theta)$ where the penalty function $p(\theta) = b'\Sigma^{-1}(\theta)b/2$. In this case we use the ordinary PL for assessment, but with an effective degrees of freedom which lies somewhere between the total number of coefficients (β, b) of 38 and the degrees of freedom for the fixed effects of 1. For this particular problem $\Sigma(\theta) = \theta I$, a multiple of the identity matrix and our penalty is $p(\theta) = -1/(2\theta) \sum j = 137 b_j^2$. As the penalty goes to zero the fit will be equivalent to treating factor(center) as a fixed effect with degrees of freedom equal to 38, as the penalty becomes larger the effective degrees of freedom will decrease. In this case a test for the random effects would be based on (319.7 - 105.7) on 27.7 degrees of freedom. We do not normally use this test, but the adjusted degrees of freedom is useful in summarizing the fit.

The major components of the model can be extracted via the following functions:

	Method		
	lme/lmer	coxme < 2.2	coxme ≥ 2.2
fixed effects coefficients	fixef	fixef	fixef
random effects coefficients	ranef		ranef
fixed effects variance matrix	vcov		vcov
random effects parameters (θ)	VarCorr	ranef	VarCorr

Note that for version 2.2 and greater these follow the same pattern as the linear mixed effects models in R; prior to version 2.2 the ranef method was incorrectly assigned.

One feature of the mixed effects Cox model is that the standard deviation of the random effect is directly interpretable. The random effects b_j for each center j are in the risk score, a value of .33 for instance (one standard deviation above the mean) corresponds to a relative risk of $\exp(.33) = 1.39$, an almost 40% higher risk of death for subjects at that center. (In

real data sets we would of course have adjusted for imbalances in patient risk factors between centers.) The random effects coefficients can be retrieved using the `ranef` function. The code below shows center effects ranging from less than 1/2 to over 1.6 times the average risk for the study. Because there may be multiple random effects in a model the ranef function returns a list, with one element per random effect. We are interested in the first element (only element in this case) of the list.

```
> stem(exp(ranef(efit2)[[1]]))

  The decimal point is 1 digit(s) to the left of the |

   4 | 44
   6 | 42447
   8 | 679015556888
  10 | 12458966
  12 | 45670
  14 | 159
  16 | 27
```

To look at random treatment effects within center we can add a nested effect

```
> efit3 <- coxme(Surv(y, uncens) ~ trt + (1 | center/trt), eortc)
> efit3

Cox mixed-effects model fit by maximum likelihood
  Data: eortc
  events, n = 1463, 2323
  Iterations= 10 54
                     NULL Integrated     Fitted
Log-likelihood -10638.71  -10517.57 -10464.38

                    Chisq    df p    AIC    BIC
Integrated loglik 242.28  3.00 0 236.28 220.42
 Penalized loglik 348.67 39.26 0 270.16  62.56

Model:  Surv(y, uncens) ~ trt + (1 | center/trt)
Fixed coefficients
         coef exp(coef)   se(coef)    z p
trt 0.7420388  2.100213 0.08270483 8.97 0

Random effects
 Group      Variable    Std Dev    Variance
 center/trt (Intercept) 0.20451052 0.04182455
 center     (Intercept) 0.26273062 0.06902738
```

This shows a further improvement in fit, but by much smaller amount.

There is one important difference between how random effects in coxme and in lmer are handled concerning intercepts. Linear models have an implied intercept term. Unless explicitly

removed, the model `lm(y ~x)` will actually fit the model `lm(y ~x +1)`. This carries forward into the notation of lmer where a random effect `(age | group)` will automatically add the random intercept and fit `(1 + age | group)`. Cox models do not have an intercept term, and an automatic "1" is not added to either fixed effects *or* random effects portions of the formula.

4 The Minnesota Breast Cancer Family Study

4.1 Background

Details of the study can be found in Sellers [4] Briefly, in 1944 a family study of breast cancer was initiated at the Dight institute for Human Genetics at the University of Minnesota to examine the influence of heredity on the risk of breast cancer. Probands were a consecutive series of all breast cancer patients ascertained at the tumor clinic of the University hospital between 1944 and 1952. A total of 544 families were studied, representing data on 4418 family members. Results of that early study provided some of the first evidence that breast cancer clusters in families [1].

Records of the research then sat dormant for 40 years, until a series of follow-up studies was initiated in the late 1990s. Prevalent cases ($n = 40$) and families where most of the relatives other than the proband were deceased at baseline ($n = 42$) were excluded. Of the remainder 30 families could be contacted and 6 refused, leaving 426 participating families. After extending pedigrees to the current generation, the final data set has 28081 subjects. The full set of subjects along with a few selected variables is included in the `kinship2` package as the `minnbreast` data set. The analysis of this data set was the original genesis of the coxme function.

```
> library(coxme)
> library(kinship2)
> options(show.signif.stars=FALSE)
> makefig <- function(file) {
+     pdf(paste(file, "pdf", sep='.'), width=7, height=5)
+     par(mar=c(5.1, 4.1, .1, .1))
+ }
> names(minnbreast)

 [1] "id"        "proband"  "fatherid" "motherid" "famid"    "endage"
 [7] "cancer"    "yob"      "education" "marstat"  "everpreg" "parity"
[13] "nbreast"   "sex"      "bcpc"

> with(minnbreast, table(sex, cancer, exclude=NULL))

      cancer
sex        0      1   <NA>
  F    10618   1224    976
  M     8534    141   4827
  <NA>     4     11   1746

> mped <- with(minnbreast, pedigree(id, fatherid, motherid, sex,
+                                    affected=cancer, famid=famid,
+                                    status=proband))
```
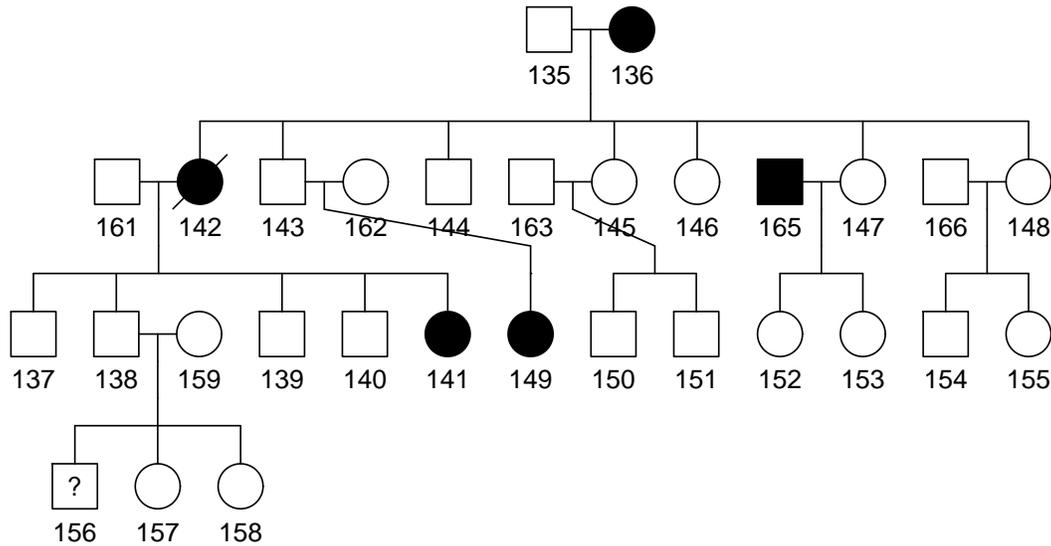
Figure 1: Pedigree for family 8 of the Minnesota Breast Cancer Study. Cancer cases are filled, the proband is marked with a slash

```
> makefig("cfig1")
> plot(mped["8"])   #figure 1

Did not plot the following people: 160 164 167 168 169 170 171 172 173 174

> dev.off()

null device
          1
```

Figure 1 shows the pedigree for family 8. The original case (the proband) has a mother, daughter, and niece with breast cancer. This is a high risk family. There is also a brother-in-law with prostate cancer.

An aside. Our first action was to turn off one of the most egregious abominations of statistical packages: using stars to train our users that "< .05" is all that matters, when in fact that is one of the most useless ways known to employ statistical methods. I also like to place my figures within the document using the latex \begin{figure} command, hence I use fig=FALSE as a default Sweave option and generate the pdfs directly.

4.2 Simple models

The simplest model is to look for a random intercept per family.

```
> minnfemale <- minnbreast[minnbreast$sex == 'F' & !is.na(minnbreast$sex),]
> fit0 <- coxph(Surv(endage, cancer) ~ I(parity>0), minnfemale,
+               subset=(proband==0))
> summary(fit0)

Call:
coxph(formula = Surv(endage, cancer) ~ I(parity > 0), data = minnfemale,
    subset = (proband == 0))

  n= 9421, number of events= 782
   (2971 observations deleted due to missingness)

                     coef exp(coef) se(coef)      z Pr(>|z|)
I(parity > 0)TRUE -0.3366    0.7142   0.1037 -3.246  0.00117


                  exp(coef) exp(-coef) lower .95 upper .95
I(parity > 0)TRUE    0.7142        1.4    0.5829    0.8752


Concordance= 0.519  (se = 0.006 )
Rsquare= 0.001   (max possible= 0.758 )
Likelihood ratio test= 9.71  on 1 df,   p=0.001831
Wald test            = 10.54  on 1 df,   p=0.001171
Score (logrank) test = 10.64  on 1 df,   p=0.001109

> fit1 <- coxme(Surv(endage, cancer) ~ I(parity>0) + (1|famid),
+               minnfemale, subset=(proband==0))
> print(fit1)

Cox mixed-effects model fit by maximum likelihood
  Data: minnfemale;  Subset: (proband == 0)
  events, n = 782, 9421 (2971 observations deleted due to missingness)
  Iterations= 17 89
                    NULL Integrated     Fitted
Log-likelihood -6690.462  -6676.827 -6576.373


                  Chisq    df          p   AIC      BIC
Integrated loglik 27.27  2.00 1.1973e-06 23.27    13.95
 Penalized loglik 228.18 93.14 2.4769e-13 41.89 -392.32

Model:  Surv(endage, cancer) ~ I(parity > 0) + (1 | famid)
Fixed coefficients
                        coef exp(coef)  se(coef)     z     p
I(parity > 0)TRUE -0.3437546 0.7091029 0.1048988 -3.28 0.001
```

```
Random effects
 Group Variable  Std Dev   Variance
 famid Intercept 0.4118759 0.1696417
```

Note that we don't include the proband in the model. Because they are the index case that caused a family to be included, they have in a probability of 1 for cancer and don't fit into the predictive framework.

From the simple Cox model we see that ever having a child is protective, reducing the risk of breast cancer by about 30%. A mixed effects model has nearly the same estimate for parity. The random family effect, an estimated intercept (excess risk) for each family, has a standard deviation of .41. We would expect about 15% of the families to be 1 standard deviation or more above the mean, and these families will have a breast cancer risk that is $\exp(.41)=1.5$ times the norm. A similar fraction have lower risk. This is a modestly large familial effect.

The result of `fit1` is also known as a *shared frailty* model. The term originally arose in demography, where the excess risk $f_k = \exp(b_k)$ for family k could be thought of as an underlying propensity for failure or "frailty" for the subject. The statistics literature is divided on the use of f or frailty as the underlying random variable or using b. The first leads to investigation of random effects distributions that are > 0 such as the gamma, positive stable, and log-normal. We strongly prefer the latter form, however, first since it ties into the very familiar notational structure of linear mixed effects models, and second because the Gaussian forms a simpler framework for multiple and correlated random effects.

```
> ncancer <- with(minnfemale, tapply(cancer, famid, sum, na.rm=T))
> pyears <-  with(minnfemale, tapply(endage -18, famid, sum, na.rm=T))
> count  <-  with(minnfemale, tapply(cancer, famid,
+                                    function(x) sum(!is.na(x))))
> indx <- match(names(ranef(fit1)[[1]]), names(ncancer))
> makefig("cfig2")
> plot(ncancer[indx], exp(ranef(fit1)[[1]]), log='y',
+      xlab="Number of cancers per family",
+      ylab="Estimated familial risk")
> abline(h=1, lty=2)
> text(c(8.1, 1.6), c(.85, 1.2), c("165", "72"))
> dev.off()

null device
          1

> indx <- match(c(72,165), names(ncancer))
> temp <- cbind(ncancer, count, pyears, 100*ncancer/pyears)[indx,]
> dimnames(temp) <- list(c(72, 165),
+                        c("Cancers", "N", "Years of FU", "Rate"))
> print(round(temp,2))

    Cancers  N Years of FU Rate
72        2  5      109.50 1.83
165       8 134     5444.21 0.15
```
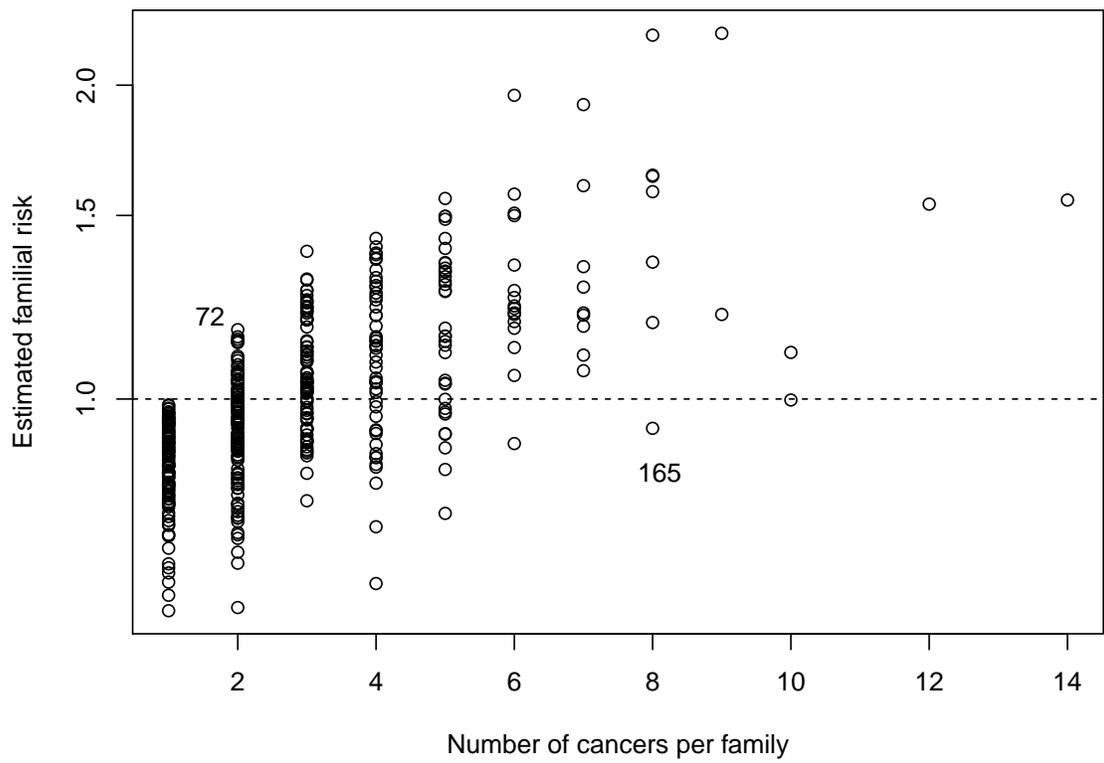
9

Figure 2: The number of breast cancers in each family versus the estimated familial risk
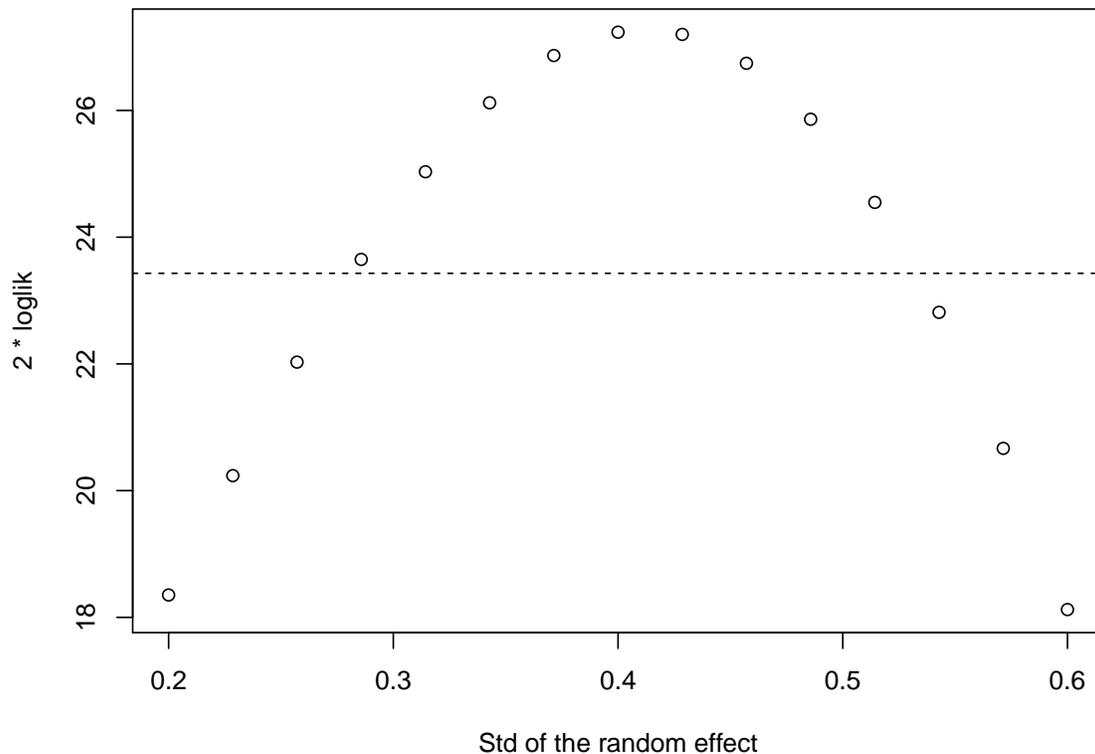
Figure 3: Profile likelihood for the per-family random effects model

Figure 2 shows number of cancers in each family versus the estimated excess risk from the model. Family 72 has a high risk but only 2 cases; there are only 5 female relatives in the family and the cancers occur at ages 32 and 36. Because the family is small the model has shrunk the estimated risk closer to the overall mean. Family 165 has 8 cancers but is has lower risk than the average Minnesota subject due to the large family size.

There is not a good closed form formula for the variance of the variance estimates themselves, i.e., one that is both reliable and simple to compute. Valid confidence intervals can be obtained from a profile likelihood, however. We start by computing the likelihood over a range of values for the variance.

```
> estvar <- seq(.2, .6, length=15)^2  #range of std values
> loglik <- double(15)
> for (i in 1:15) {
+     tfit <- coxme(Surv(endage, cancer) ~ I(parity>0) + (1|famid),
+                   data=minnfemale, subset=(proband==0),
+                   vfixed=estvar[i])
+     loglik[i] <- 2*diff(tfit$loglik)[1]
+ }
```

```
> makefig("cfig3")
> plot(sqrt(estvar), loglik,
+       xlab="Std of the random effect", ylab="2 * loglik")
> abline(h=2*diff(fit1$loglik)[1] - qchisq(.95, 1), lty=2)
> dev.off()

null device
          1
```

The 95% profile likelihood confidence interval is the region where the curve lies above the line, i.e., the set of values x for which a 1 degree of freedom likelihood ratio test would not reject the hypothesis that the true standard deviation $= x$. (It will sometimes take a couple of tries to guess an appropriate range for the variance, for many data sets it will be considerably wider than this one). We can easily get a numeric estimate from the graph, giving a confidence interval from .28 to .53.

```
> temp <- 2*diff(fit1$loglik)[1] - loglik
> approx(temp[1:8], sqrt(estvar[1:8]), 3.84)$y

[1] 0.2818699

> approx(temp[9:15], sqrt(estvar[9:15]), 3.84)$y

[1] 0.5326992
```

4.3   Correlated Random Effects

The simple family intercept model has provided some insight, but it has two serious flaws.

- Subjects will be counted in a family's risk who should actually be excluded. In family 8 there are also 4 women who have married into the family; they are not shown on figure 1 since they have no children and hence have no *genetic* relationship to family 8 at all. Yet they are counted in the familial risk.

- More generally, the simple model cannot account for degrees of association. In figure 1 subject 162 is a marry-in and so does not share the familial burden of the proband/mother/daughter cancers. Yet she is connected indirectly to the family through her high risk daughter. The largest families have over 200 individuals, some of whom are only distantly related while others are close.

We approach this by making use of the kinship matrix K. Formally, K_{ij} is the probability that for any given gene, an allele randomly chosen from subject i and another randomly chosen from subject j will be identical by descent, that is, have been passed down from a common ancestor. Then $2K$ is a measure of the expected fraction of shared genes: 1 on the diagonal, 1/2 for mother/daughter or sib/sib, 1/4 for grandparent/grandchild or uncle/niece, etc. If subject i and j are from different families then $K_{ij} = 0$ by definition, since they have no common ancestor. The relevant Cox model is

$$\lambda_i(t) = \lambda_0(t)e^{X\beta + b_i}$$
$$b_i \sim N(0, \sigma^2 2K)$$

Each subject has an individual random genetic risk b_i, but those risks are correlated according to the strength of relationship. The model is easy to fit in `coxme`

```
> kmat <- kinship(mped)
> fit2 <- coxme(Surv(endage, cancer) ~ I(parity>0) + (1|id),
+               data=minnfemale, varlist=coxmeMlist(2*kmat, rescale=F),
+               subset=(proband==0))
> print(fit2)

Cox mixed-effects model fit by maximum likelihood
  Data: minnfemale;  Subset: (proband == 0)
  events, n = 782, 9421 (2971 observations deleted due to missingness)
  Iterations= 9 49
                   NULL Integrated    Fitted
Log-likelihood -6690.462  -6671.391 -6102.548


                  Chisq     df          p   AIC      BIC
Integrated loglik  38.14   2.00 5.2159e-09 34.14    24.82
 Penalized loglik 1175.83 553.23 0.0000e+00 69.36 -2509.73


Model:  Surv(endage, cancer) ~ I(parity > 0) + (1 | id)
Fixed coefficients
                         coef exp(coef) se(coef)     z     p
I(parity > 0)TRUE -0.3602322 0.6975144 0.109819 -3.28 0.001

Random effects
 Group Variable Std Dev   Variance
 id    Vmat.1   0.8995394 0.8091712
```

The estimated genetic effect is much larger, with a standard deviation of almost 0.9. This means that there are multiple subjects in the study with quite large relative risks, $\exp(.9)= 2.5$ fold greater than the average Minnesotan.

Note that even though we are only interested in fitting the females, creation of the kinship matrix `kmat` requires use of all the subjects in the pedigree. When the model is fit appropriate rows/columns of the matrix are selected by matching its dimnames with the id variable. In fitting the model we have to explicitly describe the variance structure using one of a small set of variance functions. The coxmeMlist function accepts a set of matrices V_1, V_2, ... as arguments and fits the an overall variance matrix $\sigma_1^2 V_1 + \sigma_2^2 V_2 + \ldots$, solving for the optimal values of σ. A major portion of the code in the coxmeMlist function is devoted to ensuring that each level of the grouping variable — `id` in this instance — exists in the matrix, that all matrices have the same row and column order, and other consistency checks.

Since $K_{ij} = 0$ for any unrelated individuals, the kinship matrix is mostly zeros. Although it has a theoretical size of just under 800 million elements (`prod(dim(kmat))`) only the 500,000 non-zero elements are stored (`length(kmat@x)`) using a sparse matrix form based on the Matrix package. To maximize storage efficiency families are clustered together in adjacent rows. This carries through to the calculations, and thus it is the order of the elements in `kmat` which determines the order of the random effects coefficients b in the fitted model. Do not expect them

to be sorted in the same way as as they were in the data set, nor in the sorted order found for a factor variable in an ordinary Cox model. (For a simple model like `fit1` which does not require sparse storage they will be in standard order, however.)

4.4   Breast and prostate cancer

An interesting genetic question is whether there might be a connection between various cancers. That is, inherited traits that predispose subjects to all or some subset of malignancies. Grabrick et al [2] undertook a sub study within the Minnesota Breast Family cohort of possible genetic connections between breast and prostate cancer. A subset of 206 families was selected, and all men in those families were invited to participate. Living male members were sent a questionnaire, for those deceased or unable to answer a shortened form was sent to their spouse or close relative.

More to do for this section

5   Random slopes

References

[1] V E Anderson, H O Goodman, S C Reed. *Variables related to human breast cancer.* University of Minnesota Press, 1958.

[2] D M Grabrick, J R Cerhan, R A Veirkant, T M Therneau, J C Cheville, D J Tindall, and T A Sellers, Evaluation of familial clustering of breast and prostate cancer in the Minnesota Breast Cancer Family Study, *Cancer Detection and Prevention*, 27:30–36, 2003.

[3] S Ripatti and J Palmgren. Estimation of multivariate frailty models using penalized partial likelihood. Biometics 56:1016–1022, 2002.

[4] T A Sellers, R A King, J R Cerhan, P L Chen, D M Grabrick, L H Kushi, W S Oetting, R A Vierkant, C M Vachon, F J Couch, T M Therneau, L C Hartman and V E Anderson. Fifty year follow-up of cancer incidence in a historical cohort of Minnesota breast cancer families. *Cancer Epid Biomarkers Prev* 8:1051–7, 1999.

[5] T M Therneau and P M Grambsch, Modeling Surivival Data: Extending the Cox Model, Springer-Verlag, 2000.