

# ICPSR 2017 “Advanced Maximum Likelihood”: Survival Analysis

## Day Nine

August 17, 2017

# “Frailty” Models

$$h_i(t) = \lambda_i(t)\nu_i$$

- $\nu_i = 1 \approx$  “baseline,”
- $\nu_i > 1 \rightarrow i$  has a greater-than-average hazard,
- $\nu_i < 1 \rightarrow$  the opposite.

# More Frailty

Implies:

$$\begin{aligned} S(t|\nu_i) &= \exp \left[ - \int_0^t h(t|\nu_i) dt \right] \\ &= \exp \left[ - \int_0^t \nu_i h(t) dt \right] \\ &= \exp \left[ - \int_0^t h(t) dt \right]^{\nu_i} \\ &= S(t)^{\nu_i} \end{aligned}$$

Typically:

- Assume  $\nu_i \sim g(\nu)$ , with
- $E(\nu) = 1$  and
- $\text{Var}(\nu) = \theta$

# Example: Cox with Frailty

$$\begin{aligned}h_i(t) &= h_0(t)\nu_i\exp(\mathbf{X}_i\beta) \\ &= h_0(t)\exp(\mathbf{X}_i\beta + \alpha_i)\end{aligned}$$

where  $\alpha_i = \ln(\nu_i)$ .

(Also weibull, log-normal, etc.)

# Frailty Distributions: Gamma

$$\begin{aligned} g(\nu) &= \mathcal{G}(\theta, 1/\theta) \\ &= \frac{\nu^{1/\theta-1} \exp\left(\frac{-\nu}{\theta}\right)}{\theta^{(1/\theta)} \Gamma(1/\theta)} \end{aligned}$$

with

$$S_{\theta}(t) = \{1 - \theta \ln[S(t)]\}^{-1/\theta}$$

# Frailty Distributions: Inverse-Gaussian

$$\begin{aligned}g(\nu) &= \mathcal{IG}(\theta, 1/\theta) \\&= (2\pi\theta\nu^3)^{-1/2} \exp \left[ -\frac{1}{2\theta} \left( \alpha - 2 + \frac{1}{\nu} \right) \right]\end{aligned}$$

with

$$S_{\theta}(t) = \exp \left\{ \frac{1}{\theta} \left[ 1 - (1 - 2\theta \ln\{S(t)\})^{1/2} \right] \right\}$$

# An Important Distinction

*Individual- (or Unit-) Specific Survival Function:*

$$S(t|\nu_i) = S(t)^{\nu_i}$$

*Population Average Survival Function:*

$$\overline{S(t)} = \int_0^{\infty} S(t|\nu_i)g(\nu)d\nu$$

- Originally: E-M algorithm (e.g. Klein 1992)
- Later: Penalized Likelihood
  - Two-level iterative procedure
  - Intuition: Iterate between fitting  $\hat{\beta}|\theta$  for a range of  $\theta$ s, and searching over the (univariate) marginal likelihood for  $\theta$  to obtain  $\hat{\theta}$
  - Details: Therneau and Grambsch (2000, §9.6)



# Practical Matters

- Computation...

*"...if there are 300 families, each with their own frailty, and four other variables, then the full information matrix has  $304^2 = 92,416$  elements. The Cholesky decomposition must be applied to this matrix with each Newton-Raphson iteration."*

*– Therneau and Grambsch (2000, p. 258)*

- Fitting choices (fix  $\theta$  vs. estimation, etc.)
- Predictions / interpretation (typically assume  $\hat{\nu}_i = 1$ ).

## R

- `survival`: Fits a single `frailty` term via `frailty.gamma`, `frailty.gaussian`, or `frailty.t` to either Cox or parametric models.
- `coxme` (Cox w/Gaussian random effects; see below)
- `frailtypack` (parallel to `frailty` and `coxme`)
- Others (see the [task view](#))

## Stata

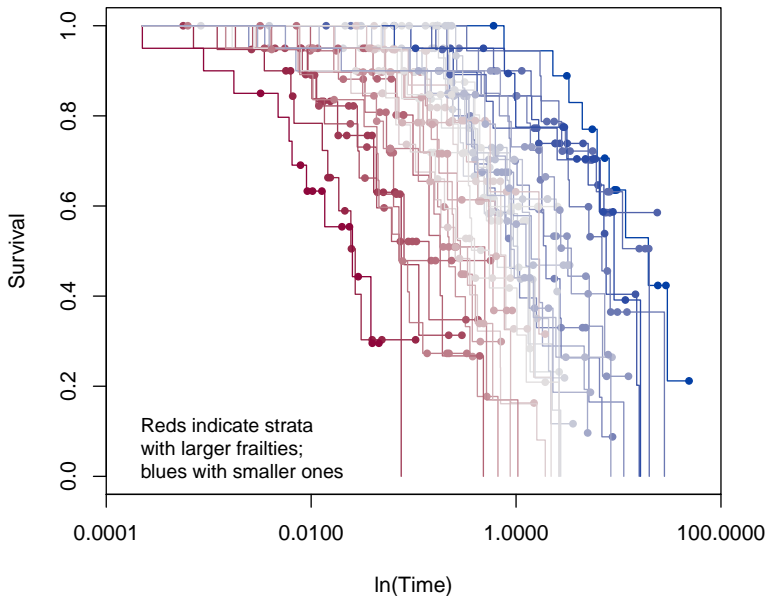
- The option `shared()` introduces one-level gamma-distributed frailties into `stcox`
- `streg` allows unshared or shared frailties (via `frailty()` and `shared()`, respectively) in both gamma and inverse-gaussian flavors in its parametric survival models; see [Guiterrez \(2002\)](#) for a good starting point.

# Simulated Example

```
> set.seed(7222009)
> G<-1:40          # "groups"
> F<-rnorm(40)     # frailties
> data<-data.frame(cbind(G,F))
> data<-data[rep(1:nrow(data),each=20),]
> data$X<-rbinom(nrow(data),1,0.5)
> data$T<-rexp(nrow(data),rate=exp(0+1*data$X+(2*data$F)))
> data$C<-rbinom(nrow(data),1,0.5)
> data<-data[order(data$F),]

> S<-Surv(data$T,data$C)
```

# K-M Plots By Strata



# Cox Fit (No Frailty)

```
> cox.noF<-coxph(S~X,data=data)
> summary(cox.noF)
Call:
coxph(formula = S ~ X, data = data)

      n= 800, number of events= 381

      coef exp(coef) se(coef)      z    Pr(>|z|)
X 0.522      1.685      0.104 5.02 0.00000051 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

      exp(coef) exp(-coef) lower .95 upper .95
X          1.69          0.593      1.37      2.07

Concordance= 0.577 (se = 0.015 )
Rsquare= 0.031 (max possible= 0.996 )
Likelihood ratio test= 25.2 on 1 df,  p=0.000000521
Wald test               = 25.2 on 1 df,  p=0.000000508
Score (logrank) test = 25.8 on 1 df,  p=0.000000382
```

# Weibull Fit (No Frailty)

```
> weib.noF<-survreg(S~X,data=data,dist="weib")  
> summary(weib.noF)
```

Call:

```
survreg(formula = S ~ X, data = data, dist = "weib")
```

	Value	Std. Error	z	p
(Intercept)	1.595	0.1450	11.00	3.92e-28
X	-1.031	0.1974	-5.22	1.76e-07
Log(scale)	0.653	0.0383	17.04	3.98e-65

Scale= 1.92

Weibull distribution

Loglik(model)= -581    Loglik(intercept only)= -594

Chisq= 27 on 1 degrees of freedom, p= 0.00000023

Number of Newton-Raphson Iterations: 5

n= 800

# Cox Fit With Frailty

```
> cox.F<-coxph(S~X+frailty.gaussian(F),data=data)
> summary(cox.F)
Call:
coxph(formula = S ~ X + frailty.gaussian(F), data = data)
```

```
n= 800, number of events= 381
```

	coef	se(coef)	se2	Chisq	DF	p
X	1.01	0.112	0.112	81.9	1.0	0
frailty.gaussian(F)				609.0	37.6	0

	exp(coef)	exp(-coef)	lower .95	upper .95
X	2.76	0.363	2.21	3.43

```
Iterations: 7 outer, 47 Newton-Raphson
```

```
Variance of random effect= 1.8
```

```
Degrees of freedom for terms= 1.0 37.6
```

```
Concordance= 0.791 (se = 0.017 )
```

```
Likelihood ratio test= 414 on 38.5 df, p=0
```

# Weibull Fit With Frailty

```
> weib.F<-survreg(S~X+frailty.gaussian(F),data=data,dist="weib")
```

```
> summary(weib.F)
```

Call:

```
survreg(formula = S ~ X + frailty.gaussian(F), data = data, dist = "weib")
```

	Value	Std. Error	z	p
(Intercept)	0.6188	0.2622	2.36	1.83e-02
X	-1.1386	0.1121	-10.16	3.12e-24
Log(scale)	0.0546	0.0417	1.31	1.91e-01

Scale= 1.06

Weibull distribution

Loglik(model)= -372    Loglik(intercept only)= -594

Chisq= 443 on 37 degrees of freedom, p= 0

Number of Newton-Raphson Iterations: 5 18

n= 800



# Example: Leader Tenure

```
> lead.S<-Surv(lead$tenstart,lead$tenure,lead$tenureend)

> Rs<-as.matrix(lead[,13:17])
> lead$region<-factor((Rs %*% 1:ncol(Rs))+1,
                      labels=c("NorthAm",colnames(Rs)))
> rm(Rs)

> lead.F<-coxph(lead.S~female*region+frailty.gamma(leadid),data=lead)
```

RStudio File Edit Code View Plots Session Build Debug Tools Window Help Tue 9:30 PM Christopher Zorn

noteOnLHR2008.R ICPSR-AdvMLE-Day9-2014.R ICPSR-AdvMLE-Day6-2014.R

```

75 lead.S<-Surv(lead$tenstart,lead$tenure,lead$tenureend)
76
77 Rs<-as.matrix(lead[,13:17])
78 lead$region<-factor((Rs %>% 1:ncol(Rs))+1,
79                     labels=c("NorthAm",colnames(Rs)))
80 rm(Rs)
81
82 lead.F<-coxph(lead.S~female*reg
83
84
85
86
87
88 pdf("lead-KM.pdf",6,5)
89

```

Environment History Presentation

Global Environment

ig num [1:701] NA NA NA NA NA NA NA N...

lead.F Large coxph.penal (31 elements, [1...

Help Viewer

Export Clear All

**R Session Aborted**

R encountered a fatal error.  
The session was terminated.

Start New Session

Console ~/Dropbox/ICPSR 2014/

```

female:regionAfrica 2.193 0
female:regionAsia 0.153 6
female:regionMidEast 0.296 3

```

Iterations: 10 outer, 83 Newton-Raphson

Variance of random effect= 0.24 I-likelihood = -19476.6

Degrees of freedom for terms= 1.0 1.3 119.3 4.9

Concordance= 0.662 (se = 0.006)

Likelihood ratio test= 858 on 127 df, p=0

```

> lead.F<-coxph(lead.S~female*region+frailty.gamma(leadid),data=lead)

```

0.0 0.2

0.001 0.010 0.100 1.000 10.000

ln(Time)

# Let's Try That Again

```
> lead.F<-coxph(lead.S~female*region+frailty.gamma(ccode),data=lead)
```

```
Warning message:
```

```
In coxpenal.fit(X, Y, strats, offset, init = init, control, weights = weights, :
```

```
  Inner loop failed to coverage for iterations 2 3
```

```
> summary(lead.F)
```

```
Call:
```

```
coxph(formula = lead.S ~ female * region + frailty.gamma(ccode),  
      data = lead)
```

```
n= 15222, number of events= 2806
```

```
(22 observations deleted due to missingness)
```

	coef	se(coef)	se2	Chisq	DF	p
female	1.2427	0.462	0.4594	7.24	1	0.007100
regionLatinAm	-0.1259	0.208	0.0333	0.37	1	0.540000
regionEurope	0.0414	0.160	0.0545	0.07	1	0.800000
regionAfrica	-0.7047	0.160	0.0840	19.45	1	0.000010
regionAsia	-0.3896	0.164	0.0742	5.65	1	0.017000
regionMidEast	-0.7478	0.186	0.0986	16.13	1	0.000059
frailty.gamma(ccode)				523.81	119	0.000000
female:regionLatinAm	-1.8826	0.851	0.8495	4.89	1	0.027000
female:regionEurope	-1.5424	0.624	0.6212	6.11	1	0.013000
female:regionAfrica	0.7854	0.861	0.8556	0.83	1	0.360000
female:regionAsia	-1.8765	0.572	0.5666	10.76	1	0.001000
female:regionMidEast	-1.2175	0.861	0.8551	2.00	1	0.160000

```
Iterations: 10 outer, 83 Newton-Raphson
```

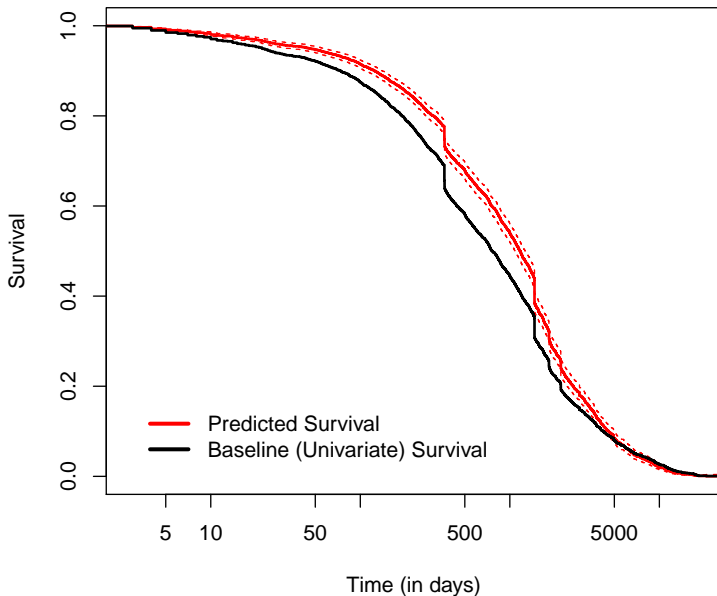
```
Variance of random effect= 0.24 I-likelihood = -19476.6
```

```
Degrees of freedom for terms= 1.0 1.3 119.3 4.9
```

```
Concordance= 0.662 (se = 0.006 )
```

```
Likelihood ratio test= 858 on 127 df, p=0
```

# Predicted vs. Actual



# Extensions: Mixed-Effects Survival Models

- HLMs for survival data / outcomes
- Combined fixed, random, and mixed effects (random-coefficient) models
- R: Implemented in `coxme`
- Stata: `stmixed` (parametric models)
- Terry Therneau has a nice [vignette](#)

# Mixed Effects Example

```
> lead.coxME<-coxme(lead.S~female + (1 | ccode/female),data=lead)
> lead.coxME
Cox mixed-effects model fit by maximum likelihood
  Data: lead
    events, n = 2806, 15222 (22 observations deleted due to missingness)
  Iterations= 38 160
                NULL Integrated Fitted
Log-likelihood -19738      -19505 -19314
```

	Chisq	df	p	AIC	BIC
Integrated loglik	465	3	0	459	441
Penalized loglik	849	129	0	590	-177

```
Model: lead.S ~ female + (1 | ccode/female)
Fixed coefficients
```

	coef	exp(coef)	se(coef)	z	p
female	-0.07	0.93	0.22	-0.31	0.75

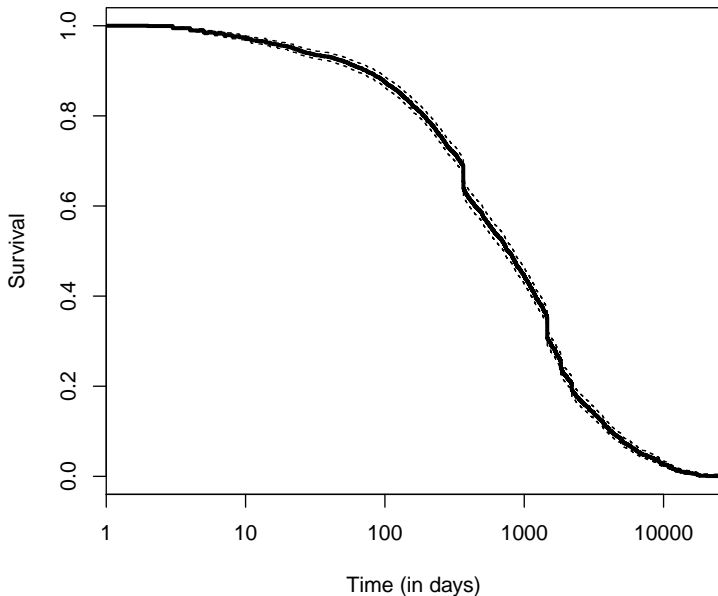
Random effects

Group	Variable	Std Dev	Variance
ccode/female	(Intercept)	0.279	0.078
ccode	(Intercept)	0.487	0.237

# Stratify? Frailties? Clustering?

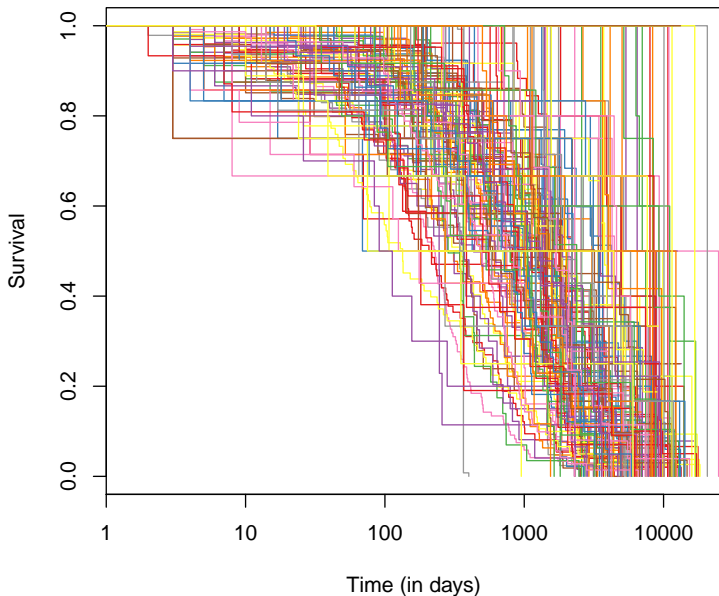
- Stratification  $\approx$  “fixed effects”
- Frailties  $\approx$  “random effects”
- “Robust” / cluster  $\approx$  GEE / PCSEs, etc.
- Not all combinations are possible, or make sense

# K-M Plot: Leaders

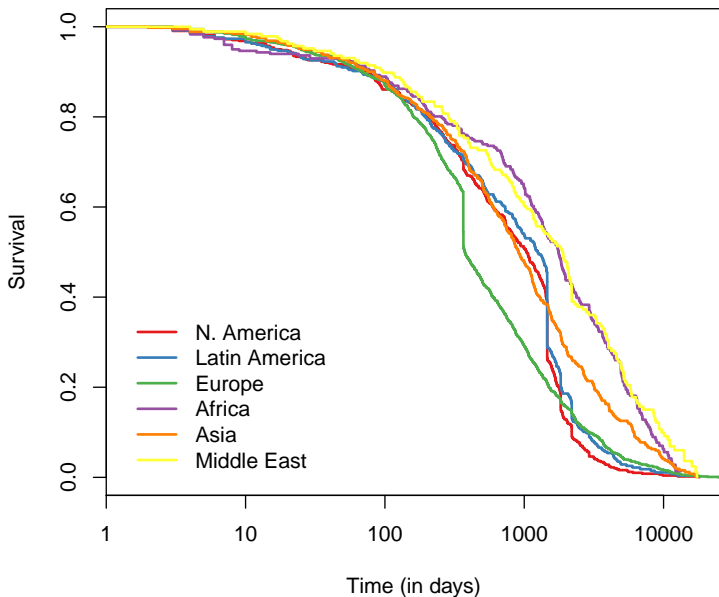




# K-M Plot: Leaders (by country)



# K-M Plot: Leaders (by region)



# Strata + Frailty

```
> lead.Fstrat<-coxph(lead.S~female*strata(region)+  
                      frailty.gamma(ccode),data=lead)
```

Warning message:

In coxpenal.fit(X, Y, strats, offset, init = init, control, weights = weights,  
 Inner loop failed to coverge for iterations 2 3 4

```
> summary(lead.Fstrat)
```

Call:

```
coxph(formula = lead.S ~ female * strata(region) + frailty.gamma(ccode),  
      data = lead)
```

n= 15222, number of events= 2806  
(22 observations deleted due to missingness)

	coef	se(coef)	se2	Chisq	DF	p
female	1.46	0.463	0.461	9.88	1	0.00170
frailty.gamma(ccode)				594.82	121	0.00000
female:strata(region)regi	-2.20	0.853	0.851	6.63	1	0.01000
female:strata(region)regi	-1.75	0.625	0.623	7.81	1	0.00520
female:strata(region)regi	0.13	0.869	0.864	0.02	1	0.88000
female:strata(region)regi	-2.07	0.573	0.568	13.04	1	0.00031
female:strata(region)regi	-1.31	0.862	0.857	2.32	1	0.13000

# Strata + Clustering

```
> lead.stratCl<-coxph(lead.S~female*strata(region)+
                      cluster(ccode),data=lead)

> summary(lead.stratCl)
Call:
coxph(formula = lead.S ~ female * strata(region) + cluster(ccode),
      data = lead)

n= 15222, number of events= 2806
(22 observations deleted due to missingness)

               coef exp(coef) se(coef) robust se      z
female               1.234    3.436    0.453    0.288  4.28
female:strata(region)region=LatinAm -1.881    0.152    0.842    0.627 -3.00
female:strata(region)region=Europe  -1.618    0.198    0.610    0.415 -3.90
female:strata(region)region=Africa   0.473    1.605    0.849    0.382  1.24
female:strata(region)region=Asia    -1.711    0.181    0.555    0.342 -5.00
female:strata(region)region=MidEast -0.709    0.492    0.846    0.349 -2.03

Concordance= 0.503 (se = 0.002 )
Rsquare= 0.001 (max possible= 0.864 )
Likelihood ratio test= 13.8 on 6 df,  p=0.0323
Wald test               = 81.6 on 6 df,  p=1.67e-15
Score (logrank) test = 20.1 on 6 df,  p=0.00263,  Robust = 14.4 p=0.0255
```

(Note: the likelihood ratio and score tests assume independence of observations within a cluster, the Wald and robust score tests do not).

From the frailty documentation:

“Note that use of a frailty term implies a mixed effects model and use of a cluster term implies a GEE approach; these cannot be mixed.”

```
> lead.FstratCl<-coxph(lead.S~female*strata(region)+frailty.gamma(ccode)+
                        cluster(ccode),data=lead)
Error in residuals.coxph(fit2, type = "dfbeta", collapse = cluster,
weighted = TRUE) :
  length of 'dimnames' [2] not equal to array extent
In addition: Warning message:
In coxpenal.fit(X, Y, strats, offset, init = init, control, weights = weights,
  Inner loop failed to coverge for iterations 2 3 4
```



693 posts

In reply to [this post](#) by Ehsan Karim

Addition of a `cluster()` term fits a Generalized Estimating Equations (GEE) type of model, addition of `frailty()` fits a random effects model (Mixed Effect or ME). In glm analysis (linear regression, logistic regression, etc) the arguments about the advantages/disadvantages of GEE vs ME would easily fill a volume. Most of this argument carries over to the coxph case; I find both approaches useful.

Caveats:

1. Coxph with `cluster()` only allows the "working independence" variance structure. The details for other variance structures were worked out by Alicia Z in her Iowa State PhD thesis, but I've never gotten around to implementing it.
2. For random effects, the `coxme` function is preferred.
3. In comparing GEE and ME one part of the argument is that the former model is "marginal" and the second "conditional", and thus the coefficients from the models mean different things. I take this with a grain of salt. Remember that ALL models are wrong.

Terry Therneau

---

[\[hidden email\]](#) mailing list

<https://stat.ethz.ch/mailman/listinfo/r-help>

PLEASE do read the posting guide <http://www.R-project.org/posting-guide.html> and provide commented, minimal, self-contained, reproducible code.

# Topics We Didn't Cover

- ★ Joint Models for Survival and Longitudinal Outcomes
  - e.g., survival + binary / multinomial / continuous variables
  - *inter alia* R package JM (Rizopolous 2010)
  - Recent reference is Viviani et al. (2014)
- ★ Causal Inference (IVs, RDDs, matching, etc.)
- ★ Variable Selection: regularization, bagging, boosting, stacking, lasso, etc.
- ★ Bayesian approaches (esp. for high-dimensional competing risks & hierarchical models); see Ibrahim et al. (2005)
- ★ New / better tools for interpretation and graphics (e.g. simPH)

# General Tips

## Journals:

- *Biometrics / Biometrika*
- *Statistics in Medicine*
- *Statistical Methods in Medical Research*
- *Lifetime Data Analysis*

## Places:

- Biostatistics / Epidemiology / Public Health
- Statistics departments
- *Not* economics, psychology, etc.



# Final Announcements

- UMICH Uniqnames expire 8/19/2017
- Contact info (8/19/2017 - )
  - Zorn: [zorn@psu.edu](mailto:zorn@psu.edu)
  - Lewis: [lewis.dom.h@gmail.com](mailto:lewis.dom.h@gmail.com)
- Github repo will be public through 9/30/2017, then go dark.