Model Assessment and Comparison Examples

Lin Zhang

Department of Biostatistics
School of Public Health
University of Minnesota

Cross-Study (Meta-analysis) Data

- Data: Estimated log relative hazards $Y_{ij} = \hat{\beta}_{ij}$ obtained by fitting separate Cox proportional hazards regressions to the data from each of $J = 18$ clinical units participating in $I = 6$ different AIDS studies.

- To these data we wish to fit the cross-study model,

$$Y_{ij} = a_i + b_j + s_{ij} + \epsilon_{ij}, \; i = 1, \ldots, I, \; j = 1, \ldots, J,$$

$$\begin{array}{rcl} \text{where } a_i & = & \text{study main effect} \\ b_j & = & \text{unit main effect} \\ s_{ij} & = & \text{study-unit interaction term, and} \\ \epsilon_{ij} & \overset{iid}{\sim} & N(0, \sigma_{ij}^2) \end{array}$$

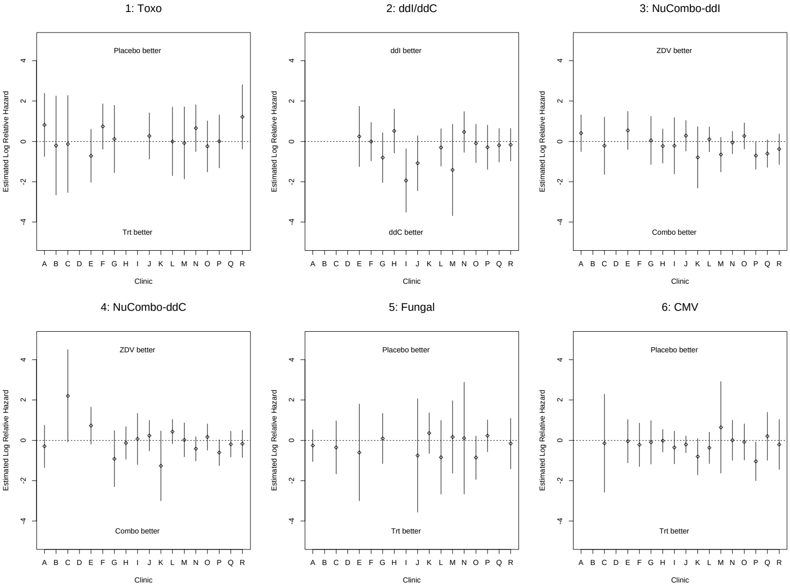
  and the estimated standard errors from the Cox regressions are used as (known) values of the σ_{ij}.

Cross-Study (Meta-analysis) Data

Estimated Unit-Specific Log Relative Hazards

	Toxo	ddI/ddC	NuCombo	NuCombo	Fungal	CMV
Unit			ZDV+ddI	ZDV+ddC		
A	0.814	NA	-0.406	0.298	0.094	NA
B	-0.203	NA	NA	NA	NA	NA
C	-0.133	NA	0.218	-2.206	0.435	0.145
D	NA	NA	NA	NA	NA	NA
E	-0.715	-0.242	-0.544	-0.731	0.600	0.041
F	0.739	0.009	NA	NA	NA	0.222
G	0.118	0.807	-0.047	0.913	-0.091	0.099
H	NA	-0.511	0.233	0.131	NA	0.017
I	NA	1.939	0.218	-0.066	NA	0.355
J	0.271	1.079	-0.277	-0.232	0.752	0.203
K	NA	NA	0.792	1.264	-0.357	0.807
L	-0.002	0.300	-0.103	-0.431	0.837	0.373
M	-0.076	1.413	0.658	-0.022	-0.164	-0.64
N	0.651	-0.470	0.060	0.421	-0.112	-0.010
O	-0.249	0.098	-0.272	-0.163	0.860	0.081
P	0.003	0.292	0.705	0.608	-0.327	1.044
Q	NA	0.195	0.605	0.187	NA	-0.201
R	1.217	0.165	0.385	0.172	-0.022	0.203

Cross-Study (Meta-analysis) Data



1: Toxo

2: ddI/ddC

3: NuCombo-ddI

4: NuCombo-ddC

5: Fungal

6: CMV

Cross-Study (Meta-analysis) Data

- *Goal:* To obtain fitted y_{ij}, and identify which clinics are opinion leaders (strongly agree with overall result across studies) and which are dissenters (strongly disagree).

- Here, overall results all favor the treatment (i.e. mostly negative Ys) except in Trial 1 (Toxo). Thus we multiply all the Y_{ij}'s by -1 for $i \neq 1$, so that larger Y_{ij} correspond to stronger agreement with the overall in all cases.

- Note that some values are missing ("NA") since
  - not all 18 units participated in all 6 studies
  - the Cox estimation procedure did not converge for some units that had few deaths

Cross-Study (Meta-analysis) Data

- With $I + J + IJ$ parameters but fewer than IJ data points, some effects must be treated as random!

- **Second stage** of our model:

  $$a_i \overset{iid}{\sim} N(0, 100^2), \quad b_j \overset{iid}{\sim} N(0, \sigma_b^2), \quad \text{and} \quad s_{ij} \overset{iid}{\sim} N(0, \sigma_s^2)$$

  Third stage of our model:

  $$\sigma_b \sim Unif(0.01, 100) \quad \text{and} \quad \sigma_s \sim Unif(0.01, 100)$$

  That is, we

  - preclude borrowing of strength across studies, but
  - encourage borrowing of strength across units

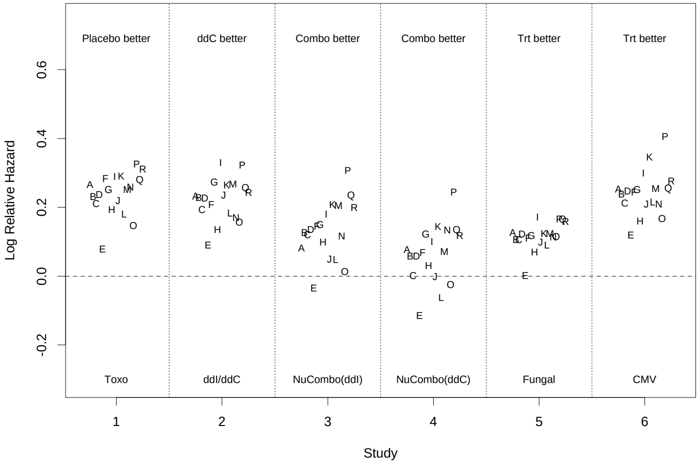
- WinBUGS code to do the analysis ...

Questions

- Which clinical unit has the most positive effect?

- Which clinical unit has the most negative effect?

Plot of posterior means of $\theta_{ij} = a_i + b_j + s_{ij}$



- ◇ Unit P ($j = 16$) is an opinion leader; Unit E ($j = 5$) is a dissenter
- ◇ Substantial shrinkage towards 0 has occurred: mostly positive values; no estimated θ_{ij} greater than 0.6

Model Assessment

- We assess the overall fitness of the two-way ANOVA model using the Bayesian p-value:

  $$\begin{aligned} p_{post} &= E_{\boldsymbol{\theta}|\mathbf{y}}[P(T(\mathbf{y}^{rep}, \boldsymbol{\theta}) > T(\mathbf{y}, \boldsymbol{\theta}))] \\ &= \int P(T(\mathbf{y}^{re}, \boldsymbol{\theta}) > T(\mathbf{y}, \boldsymbol{\theta})|\boldsymbol{\theta})\ p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}, \\ &\approx \frac{1}{G}\sum_{g=1}^{G} I_{T(\mathbf{y}^{rep(g)}, \boldsymbol{\theta}^{(g)}) > T(\mathbf{y}, \boldsymbol{\theta}^{(g)})} \end{aligned}$$

  where $\boldsymbol{\theta}^{(g)} \sim p(\boldsymbol{\theta}|\mathbf{y})$ and then $\mathbf{y}^{rep(g)} \sim p(\mathbf{y}^{rep}|\boldsymbol{\theta}^{(g)})$.

- Here we consider test quantity / "discrepancy measure" to be the sum of squared standardized residuals

  $$T(\mathbf{y}, \boldsymbol{\theta}) = \sum_{i,j} \frac{(y_{ij} - \theta_{ij})^2}{Var(y_{ij})}$$

  Good for "omnibus goodness-of-fit" measure

Marginal Check

- Conduct marginal check using the marginal p-value:

$$\begin{aligned} p_{ij} &= E_{\boldsymbol{\theta}|\mathbf{y}}[P(T(y_{ij}^{rep},\boldsymbol{\theta}) > T(y_{ij},\boldsymbol{\theta}))] \\ &= \int P(T(y_{ij}^{rep},\boldsymbol{\theta}) > T(y_{ij},\boldsymbol{\theta})|\boldsymbol{\theta})\; p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}, \\ &\approx \frac{1}{G}\sum_{g=1}^{G} I_{T(y_{ij}^{rep(g)},\boldsymbol{\theta}^{(g)})>T(y_{ij},\boldsymbol{\theta}^{(g)})} \end{aligned}$$

  where $\boldsymbol{\theta}^{(g)} \sim p(\boldsymbol{\theta}|\mathbf{y})$ and then $\mathbf{y}^{rep(g)} \sim p(\mathbf{y}^{rep}|\boldsymbol{\theta}^{(g)})$.

- Consider the same test quantity: squared standardized residuals

$$T(y_{ij},\boldsymbol{\theta}) = \frac{(y_{ij}-\theta_{ij})^2}{Var(y_{ij})}$$

Questions

- Does the Bayesian p-value suggest an overall model fitness?
- Which data point is likely to be an outlier based on marginal p-values?

Model Comparison

- Since we lack replications for each study-unit (i-j) combination, the interactions s_{ij} in this model were only weakly identified, and the model might well be better off without them (or even without the unit effects b_j). As such, compare a variety of reduced models:

```
  Y[i,j] ~ dnorm(theta[i,j],P[i,j])
#M1:    theta[i,j] <- a[i]+b[j]+s[i,j]  # full model
#M2:    theta[i,j] <- a[i] + b[j]       # drop interactions
#M3:    theta[i,j] <- a[i]              # study effect only
#M4:    theta[i,j] <- b[j]              # unit effect only
```

- We use DIC and WAIC to compare these four models.

DIC and WAIC calculation in WinBUGS

- DIC is directly available in WinBUGS.
- WAIC is given by

$$\begin{aligned} WAIC &= -2lppd + 2p_{WAIC} \\ &= -2\sum_{i=1}^{n} \log\left(E_{\theta|\mathbf{y}}\left[p(y_i|\theta)\right]\right) + 4\sum_{i=1}^{n} var_{\theta|\mathbf{y}}[\log p(y_i|\theta)] \end{aligned}$$

  - Each term can be computed via Monte Carlo methods given posterior samples of point predictive density $p(y_i|\theta)$ and log point predictive density $\log p(y_i|\theta)$

Questions

- Which model do you choose based on DIC?

- Which model do you choose based on WAIC?

DIC results for Cross-Study Data:

model	$\overline{D}$	p_D	DIC
full model	122.0	12.8	134.8
drop interactions	123.4	9.7	133.1
study effect only	126.0	6.0	132.0
unit effect only	122.9	6.2	129.1

- The DIC-best model is the one with only the unit effects b_j.
- These DIC differences are not much larger than their possible Monte Carlo errors, so almost any of these models could be justified here.