Random Effect Modeling

What is it for, and why do it ?

Regression and fixed effects

- The most basic model we usually assume which relates an outcome to a predictor/covariate is a regression model:

$$y_i = \mu_i + e_i$$

where

$$\mu_i = \alpha_0$$

or if there is a predictor then

$$\mu_i = \alpha_0 + \alpha_1 x_{1i}$$

- This assumes a linear relation and a single error term

Error assumption

- The linear model above assumes variation of Y around a mean level with a single error term (e). This is true for Gaussian models for example, where e has a Gaussian distribution
- In many studies there can be extra noise in the outcome. Frailty/susceptibility is a unit level effect for example.
- Hence there could be an extra underlying effect in the relation: $y_i = \mu_i + e_{1i} + e_{2i}$
- So that Y doesn't vary directly around the mean

Variance components

- This situation used to be called 'variance components' modeling
- Now they are usually called 'random effect' models
- In Gaussian models, the effects are additive and so there must be some distinction between the effects otherwise they are not identified.
- In discrete data models (Poisson, binomial, categorical) this is not such an issue.
- However identification of effects is a concern for such models

Simple models

- Assume a simple Gaussian linear regression:

$$y_i = \mu_i + e_{1i} + e_{2i}$$

where

$$\mu_i = \alpha_0 + \alpha_1 x_{1i}$$

and so

$$y_i = (\alpha_0 + e_{1i}) + \alpha_1 x_{1i} + e_{2i}$$

- A random intercept model which addresses the extra noise of the outcome (frailty/susceptibility).
- If Y was observed multiple times then we would have an independent estimate of (pure) error, but if not then assumptions would have to be made to distinguish the intercept and global error.
- If you believe that extra variability exists in Y then this is how to model it.

Discrete outcome models

- For discrete models the random effect is less confounded. eg

$$\text{Poisson: } y_i \sim \text{Pois}(\mu_i);$$

$$\log(\mu_i) = \alpha_0 + e_{1i}: \text{ extra Poisson variation/ over dispersion}$$

$$\mu_i = \exp(\alpha_0 + e_{1i})$$

$$\text{binomial: } y_i \sim \text{bin}(p_i, n_i);$$

$$\text{logit}(p_i) = \log(p_i/(1-p_i)) = \alpha_0 + e_{1i}: \text{extra binomial variation}$$

$$p_i = \exp(\alpha_0 + e_{1i})/(1+\exp(\alpha_0 + e_{1i}))$$

Random Intercept models

- Outcome confounding in population level studies (cross-sectional, cohort, spatial) means that random intercept models should be used in general.
- Hence, it would be important to have access to software that allows this. Bayesian software such as WinBUGS/OpenBUGS, INLA, CARBayes, Nimble all allow the use of random intercept models at the unit level.
- Note that *experimental* data often does NOT require random effect modeling eg the famous Bliss Beetle LD50 study

Interpretation

- Poisson example: $y_i \sim \text{Pois}(\exp(\alpha_0 + e_{1i}))$
- Value of Y varies around $\exp(\alpha_0)$
- If we also had predictors then we are assuming that

Y response consists of the predictor effect plus the error (random effect)

- Hence the extra noise adds into the confounded outcome
- Note that even without a linear predictor, as a function of covariates, we have a mixed effect model (GLMM)

Binomial outcome

- Count in finite population:

$$y_i \sim \text{bin}(p_i, n_i)$$
$$p_i = \exp(\alpha_0 + e_{1i}) / \{1 + \exp(\alpha_0 + e_{1i})\}$$
$$E(y_i) = n_i \exp(\alpha_0 + e_{1i}) / \{1 + \exp(\alpha_0 + e_{1i})\}$$

- The random effect adjust the probability of the outcome.
- Even for binary data extra noise will affect the probability

Fitting random intercept models I

- Random effects must be assigned a structure
- It is natural to assume that they are centered on zero and symmetric
- Also if uncorrelated (extra) variation then it is usual to assume IID

$$e_{1i} \overset{iid}{\sim} N(0, \tau_e^{-1})$$

where

τ_e is a precision

- This would be called a prior distribution in Bayesian modeling

Fitting random intercept models II

- A basic Poisson model with random intercept would then be

$$y_i \sim \text{Pois}(\mu_i)$$

$$\log(\mu_i) = \alpha_0 + e_{1i}$$

$$\alpha_0 \sim N(0, \tau_0^{-1})$$

$$e_{1i} \sim N(0, \tau_e^{-1})$$

$$\tau_* \sim \text{Ga}(a,b)$$

- This can be fitted using standard Bayesian software such as BUGS or INLA

Simulated examples

- binomial_over_dispersion.txt
- This code generates outcomes from a Bernoulli model
- With different definitions of logit(p)
- Y1 constant b0 = 0; p1 = 0.5
- Y2 intercept b0 plus RE: b1 ~N(0,3)
- Y3 intercept plus 2 covariates (x1,x2) ; x1 is trended, x2 is randomly generated; no RE
- Y4 is as Y3 but with random effect (b1) with SD = 3

Models fitted using INLA

- Set an index:
- reg<-seq(1:N)
- Example of code:

```
form5<-y4~1+x1+x2+f(reg,model="iid",param=c(1,0.5))
res5<-
inla(form5,family="binomial",data=data1,Ntrials=rep(den,N),
control.compute=list(dic=TRUE,waic=TRUE,cpo=TRUE))
#res5$summary.fixed[1]
#summary(res5)
res5$dic$dic;res5$waic$waic;res5$mlik
PML<-sum(log(res5$cpo$cpo));PML
```

Simulation results table

- I simulated a variety of binomial models with and without covariates and with and without random effect noise
- I fitted a range of models to these simulated data including :
  - Constant risk only
  - Constant risk and random intercept
  - Regression only
  - Regression and random intercept

Settings

- Binomial simulation with denominator (den = 50)
- Sample size N=100
- Random effect assumed to be b1 ~N(0,sd=3)
- Fixed intercept: b0 = 0
- Y1,….Y4<- binom(N,den,p*)
- p*=exp(LN)/{1+exp(LN)}

Settings

- LN = constant (b0) y1
- LN = b0+b1 y2
- LN = b0+b*x1+c*x2 y3 where x1 and x2 are covariates (x1 trended and x2 random Gaussian)
- LN = b0+b*x1+c*x2+b1 y4

Fitted models

Models	Sim	fitted	DIC
1	Y1	fixed intercept	549.7
2	Y2	fixed intercept	4199.2
3	Y2	fixed and random intercept	476.8
4 (6)	Y4	regression + random Intercept	441.1
5 (7)	Y4	Random intercept only	441.1
6 (8)	Y4	regression only	3468.5

Notes

- Y2 random intercept simulation and the random effect fitted model does best at describing the variation (model 3 lower DIC than model 2)
- Y4 is simulated as from a regression with random effect added. The best model in terms of DIC is the regression +random intercept OR random intercept only
- Y4 the regression only model (model 6) is much higher DIC
- REs nearly always improve fit unless there is little noise.

Finally

- This demonstrates that it is very important to consider models with REs when extra noise is suspected such as in population studies.
- Even for description: RE models can describe variation well even without regressors.
- Many standard statistical packages don't allow unit level random effect models.
- Bayesian packages (WinBUGS, OpenBUGS, JAGS, INLA, Nimble, Stan) all do allow this.