

S03 - Random effects

STAT 401 (Engineering) - Iowa State University

April 23, 2018

Regression models

For continuous Y_i , we have linear regression

$$Y_i \stackrel{ind}{\sim} N(\mu_i, \sigma^2), \quad \mu_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}$$

For binary or count with an upper maximum Y_i , we have logistic regression

$$Y_i \stackrel{ind}{\sim} \text{Bin}(n_i, \theta_i), \quad \log\left(\frac{\theta_i}{1 - \theta_i}\right) = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}$$

For count data with no upper maximum, we have Poisson regression

$$Y_i \stackrel{ind}{\sim} \text{Po}(\lambda_i), \quad \log(\lambda_i) = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}$$

But what if our observations cannot reasonably be assumed to be independent given these explanatory variables?

Random effect model

Suppose we have continuous observations Y_{ij} for individual i from group j . A random effects model (with a common variance) assumes

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}, \quad \epsilon_{ij} \stackrel{\text{ind}}{\sim} N(0, \sigma_\epsilon^2)$$

and, to make the α_j random effects, independent of ϵ_{ij} assume

$$\alpha_j \stackrel{\text{ind}}{\sim} N(0, \sigma_\alpha^2).$$

This makes observations within the group correlated since

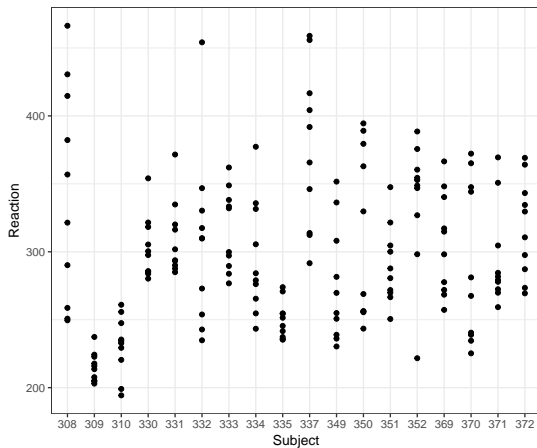
$$\begin{aligned} \text{Cov}[Y_{ij}, Y_{i'j}] &= \text{Cov}[\alpha_j + \epsilon_{ij}, \alpha_j + \epsilon_{i'j}] \\ &= \text{Var}[\alpha_j] = \sigma_\alpha^2 \end{aligned}$$

and

$$\text{Cor}[Y_{ij}, Y_{i'j}] = \frac{\text{Cov}[Y_{ij}, Y_{i'j}]}{\sqrt{\text{Var}[Y_{ij}] \text{Var}[Y_{i'j}]}} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}.$$

Sleep study example

```
ggplot(sleepstudy, aes(Subject, Reaction)) + geom_point() + theme_bw()
```



Sleep study example

```
summary(me <- lmer(Reaction ~ (1|Subject), sleepstudy))
```

Linear mixed model fit by REML ['lmerMod']

Formula: Reaction ~ (1 | Subject)

Data: sleepstudy

REML criterion at convergence: 1904.3

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.4983	-0.5501	-0.1476	0.5123	3.3446

Random effects:

Groups	Name	Variance	Std.Dev.
Subject	(Intercept)	1278	35.75
Residual		1959	44.26

Number of obs: 180, groups: Subject, 18

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	298.51	9.05	32.98

Mixed effect model

Suppose we have continuous observations Y_{ij} for individual i from group j and an associated explanatory variable X_{ij} . A mixed effect model assumes

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \alpha_j + \epsilon_{ij} \quad \epsilon_{ij} \stackrel{\text{ind}}{\sim} N(0, \sigma_\epsilon^2)$$

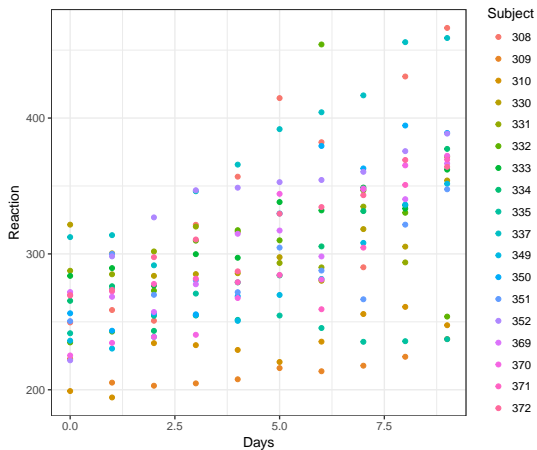
and, to make the α_i random effects, independent of ϵ_{ij}

$$\alpha_j \stackrel{\text{ind}}{\sim} N(0, \sigma_\alpha^2).$$

Again, this enforces a correlation between the observations within a group. This model is often referred to as a **random intercept model** because each group has its own intercept ($\beta_0 + \alpha_j$) and these are *random* since α_j has a distribution. Thus this model is related to a model that includes a fixed effect for each subject, but here those group specific effects are shrunk toward an overall mean (β_0).

Sleep study example

```
ggplot(sleepstudy, aes(Days, Reaction, color = Subject)) +  
  geom_point() + theme_bw()
```



Sleep study example

```
summary(me <- lmer(Reaction ~ Days + (1|Subject), sleepstudy))
```

Linear mixed model fit by REML ['lmerMod']

Formula: Reaction ~ Days + (1 | Subject)

Data: sleepstudy

REML criterion at convergence: 1786.5

Scaled residuals:

	Min	1Q	Median	3Q	Max
	-3.2257	-0.5529	0.0109	0.5188	4.2506

Random effects:

Groups	Name	Variance	Std.Dev.
Subject	(Intercept)	1378.2	37.12
Residual		960.5	30.99

Number of obs: 180, groups: Subject, 18

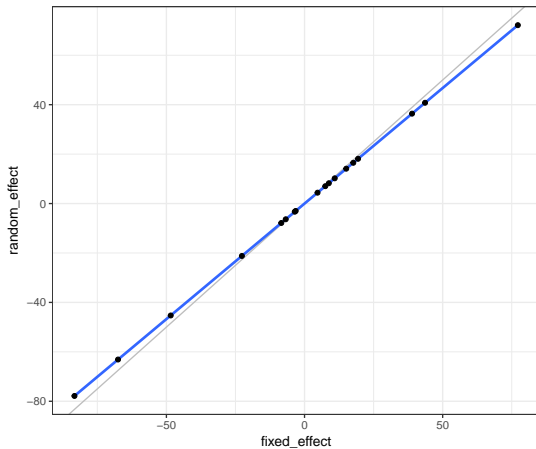
Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	251.4051	9.7467	25.79
Days	10.4673	0.8042	13.02

Correlation of Fixed Effects:

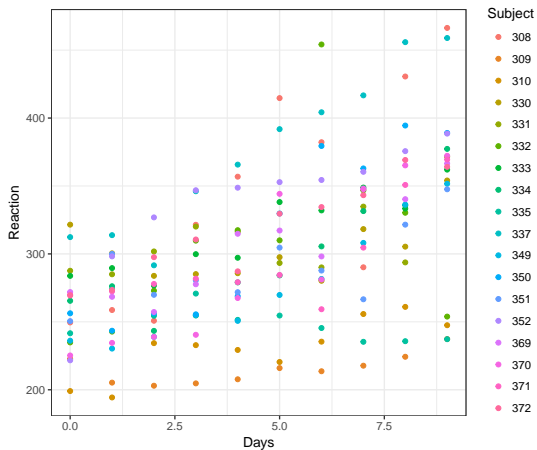
	(Intr)
Days	-0.371

Shrinkage



Sleep study example

```
ggplot(sleepstudy, aes(Days, Reaction, color = Subject)) +  
  geom_point() + theme_bw()
```



Random slope model

Suppose we have continuous observations Y_{ij} for individual i from group j . A mixed effect model with group specific slopes assumes

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \alpha_{0j} + \alpha_{1j} X_{ij} + \epsilon_{ij} \quad \epsilon_{ij} \stackrel{ind}{\sim} N(0, \sigma_\epsilon^2)$$

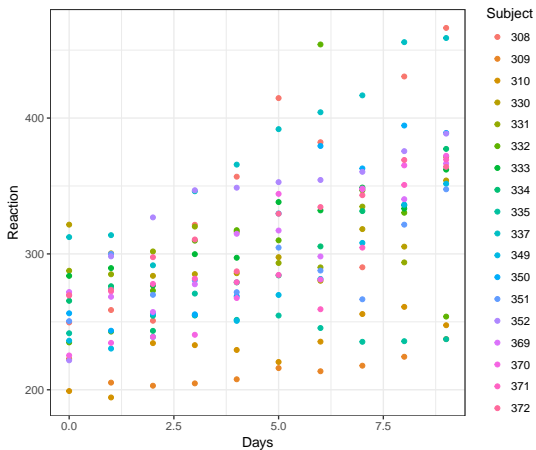
and, independent of ϵ_{ij} ,

$$\begin{pmatrix} \alpha_{0j} \\ \alpha_{1j} \end{pmatrix} \stackrel{ind}{\sim} N(0, \Sigma_\alpha)$$

$N(0, \Sigma_\alpha)$ represents a bivariate normal with mean 0 and covariance matrix Σ_α . This model is often referred to as a **random slope model** because each group has its own slope ($\beta_1 + \alpha_{1j}$) and these are *random* since α_{1j} has a distribution. Thus this model is related to a model that includes an interaction between the group and the explanatory variable, but here those subject specific slopes are shrunk toward an overall slope (β_1).

Sleep study example

```
ggplot(sleepstudy, aes(Days, Reaction, color = Subject)) +  
  geom_point() + theme_bw()
```



Sleep study example

```
summary(me <- lmer(Reaction ~ Days + (Days|Subject), sleepstudy))
```

Linear mixed model fit by REML ['lmerMod']
 Formula: Reaction ~ Days + (Days | Subject)
 Data: sleepstudy

REML criterion at convergence: 1743.6

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.9536	-0.4634	0.0231	0.4634	5.1793

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
Subject	(Intercept)	612.09	24.740	
	Days	35.07	5.922	0.07
Residual		654.94	25.592	

Number of obs: 180, groups: Subject, 18

Fixed effects:

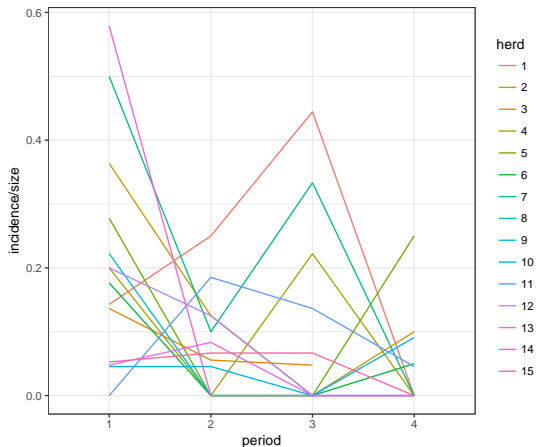
	Estimate	Std. Error	t value
(Intercept)	251.405	6.825	36.84
Days	10.467	1.546	6.77

Correlation of Fixed Effects:

	(Intr)
Days	-0.138

Contagious bovine pleuropneumonia (CBPP)

```
ggplot(cbpp, aes(period, incidence/size, color=herd, group=herd)) +  
  geom_line() + theme_bw()
```



Generalized linear mixed effect models

The same idea can be utilized in generalized linear models, e.g. logistic and Poisson regression.

A mixed effect logistic regression model for CBPP count is

$$\begin{aligned} Y_{ph} &\overset{ind}{\sim} \text{Bin}(n_{ph}, \theta_{ph}) \\ \text{logit}(\theta_{ph}) &= \beta_0 + \beta_1 \mathbf{I}(p = 2) + \beta_2 \mathbf{I}(p = 3) + \beta_3 \mathbf{I}(p = 4) + \alpha_h \\ \alpha_h &\overset{ind}{\sim} N(0, \sigma_\alpha^2) \end{aligned}$$

where $p = 1, 2, 3, 4$ stands for the period and $h = 1, \dots, 15$ stands for the herd.

When used in GLMs, these models are called generalized linear mixed models (GLMMs).

GLMMs in R

```
me <- glmer(cbind(incidence, size - incidence) ~ period + (1 | herd),
            data = cbpp, family = binomial)
summary(me)
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
Family: binomial ( logit )
Formula: cbind(incidence, size - incidence) ~ period + (1 | herd)
Data: cbpp
```

AIC	BIC	logLik	deviance	df.resid
194.1	204.2	-92.0	184.1	51

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.3816	-0.7889	-0.2026	0.5142	2.8791

Random effects:

Groups Name	Variance	Std.Dev.
herd (Intercept)	0.4123	0.6421

Number of obs: 56, groups: herd, 15

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.3983	0.2312	-6.048	1.47e-09 ***
period2	-0.9919	0.3032	-3.272	0.001068 **
period3	-1.1282	0.3228	-3.495	0.000474 ***
period4	-1.5797	0.4220	-3.743	0.000182 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Contrasts

Is there a linear trend in $\text{logit}(\theta_{ph})$ by period?

```
em <- emmeans(me, ~ period, type="response") # for interpretability
em
```

period	prob	SE	df	asyp.LCL	asyp.UCL
1	0.19807921	0.03672693	Inf	0.13569522	0.2798569
2	0.08391784	0.02363110	Inf	0.04775453	0.1433443
3	0.07401714	0.02241762	Inf	0.04040242	0.1317591
4	0.04842565	0.01959184	Inf	0.02163871	0.1048199

Confidence level used: 0.95

Intervals are back-transformed from the logit scale

```
co <- contrast(em, list(`linear trend` = c(-1.5, -0.5, 0.5, 1.5)))
confint(co)
```

contrast	odds.ratio	SE	df	asyp.LCL	asyp.UCL
linear trend	0.08735598	0.05765298	Inf	0.02396177	0.3184685

Confidence level used: 0.95

Intervals are back-transformed from the log odds ratio scale

Summary

There are a variety of opinions about when to use fixed effects and when to use random effects, e.g. [https://stats.stackexchange.com/questions/4700/](https://stats.stackexchange.com/questions/4700/what-is-the-difference-between-fixed-effect-random-effect-and-mixed-effect-mode)

[what-is-the-difference-between-fixed-effect-random-effect-and-mixed-effect-mode](https://stats.stackexchange.com/questions/4700/what-is-the-difference-between-fixed-effect-random-effect-and-mixed-effect-mode).

I am in favor of using random effects whenever we have enough levels (~ 5) of the effect to estimate the variance and we can consider the levels *exchangeable*.

For example, in the CBPP data set,

- period only has 4 levels and they are not exchangeable because they are ordered
- herd has 15 levels and the herds are exchangeable

thus I would treat period as a fixed effect and herd as random effect.