

# Linear Regression Models

## P8111

Lecture 22

Jeff Goldsmith  
April 14, 2016



THE DEPARTMENT OF  
**BIostatISTICS**



Columbia University  
MAILMAN SCHOOL  
OF PUBLIC HEALTH

# Today's Lecture

- Random intercept models ✓

- Random slope ✓

- Example (pig data!) ✓

- Example (CD4 data!) ✓

## Recall the setting

- We observe data  $y_{ij}, x_{ij}$  for subjects  $i = 1, \dots, I$  at visits  $j = 1, \dots, J_i$
- Overall, we pose the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where  $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2 V$  and

$$V = \begin{bmatrix} V_1 & 0 & \dots & 0 \\ 0 & V_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & & V_I \end{bmatrix}$$

## Recall the setting

- We've focused on random intercept models and (equivalently) uniform correlation marginal models
- Today we'll review random intercept approaches and introduce random slope models

# Random intercept model

A random intercept model with one covariate is given by

$$y_{ij} = \beta_0 + b_i + \beta_1 x_{ij} + \epsilon_{ij}$$

where

- $b_i \sim N [0, \tau^2]$
- $\epsilon_{ij} \sim N [0, \nu^2]$

# Random intercept model

More compactly, we write

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$$

where

- $\mathbf{b} \sim \text{N} [0, \tau^2 I_I]$
- $\boldsymbol{\epsilon} \sim \text{N} [0, \nu^2 I_n]$

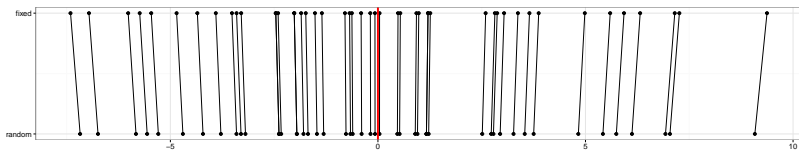
# Random intercept model

In the model

$$y = X\beta + Zb + \epsilon$$

we've discussed why we use random effects rather than fixed effects:

- Random effects induce correlation; fixed effects don't
- This reduces the number of parameters we estimate
- Random effect modeling is similar mathematically to introducing penalization



# Estimation – random intercept model

Estimation is done using MLE with model and distributional assumptions

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$$

where

- $\mathbf{b} \sim \text{N} [0, \tau^2 I_I]$
- $\boldsymbol{\epsilon} \sim \text{N} [0, \nu^2 I_n]$

Remember that BLUPs from this model can be derived without distributional assumptions (similarly to OLS and BLUEs).



## Estimation – BLUPs

Our estimate for fixed and random effects are

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{b}} \end{bmatrix} = \left( \mathbf{C}^T \mathbf{C} + \frac{\nu^2}{\tau^2} \mathbf{R} \right)^{-1} \mathbf{C}^T \mathbf{y}$$

where  $\mathbf{C} = [\mathbf{X} \ \mathbf{Z}]$  and

$$\mathbf{R} = \begin{bmatrix} 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & & 0 \\ \vdots & & \vdots & & \ddots & \\ 0 & \dots & 0 & 0 & & 1 \end{bmatrix}$$

# Random slope model

A random slope model with one covariate is given by

$$y_{ij} = \beta_0 + b_{i,0} + \beta_1 x_{ij} + b_{i,1} x_{ij} + \epsilon_{ij}$$

where

$$\begin{bmatrix} b_{i,0} \\ b_{i,1} \end{bmatrix} \sim \text{N} \left[ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_0^2 & \tau_{01} \\ \tau_{10} & \tau_1^2 \end{bmatrix} \right]$$

and

$$\epsilon_{ij} \sim \text{N} [0, \nu^2]$$

# Random slope model

Using vectors, we can write

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_0\mathbf{b}_0 + \mathbf{Z}_1\mathbf{b}_1 + \boldsymbol{\epsilon}$$

# Estimation – random slope model

Omitting the details –

- Again use MLE to set up approach and derive BLUPs (which don't depend on distributional assumptions)
- This is easier if one assumes  $\tau_{01} = 0$ , and usually the results aren't affected much
- The estimates look similar to the BLUPs for one random intercept, although there are more “ $R$ ”s to deal with
- Results again resemble ridge regression estimates, although with more than one penalty

# Random effect models

Our random slope model with one covariate given by

$$y_{ij} = \beta_0 + b_{i,0} + \beta_1 x_{ij} + b_{i,1} x_{ij} + \epsilon_{ij}$$

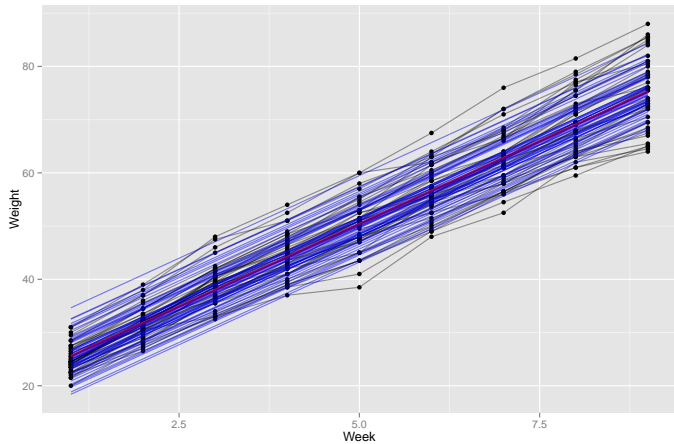
has the following properties

- $E(\mathbf{y}) = \beta_0 + \beta_1 x_{ij}$
- $E(\mathbf{y} | b_{i,0}, b_{i,1}) = (\beta_0 + b_{i,0}) + (\beta_1 + b_{i,1}) x_{ij}$

So main effect parameters are interpreted as the effect for *an average* subject; the interpretation for *a particular* subject is conditional on the random effects.

# Pig data

Random intercept model fit for pig data



# Pig data

## Random intercept model code

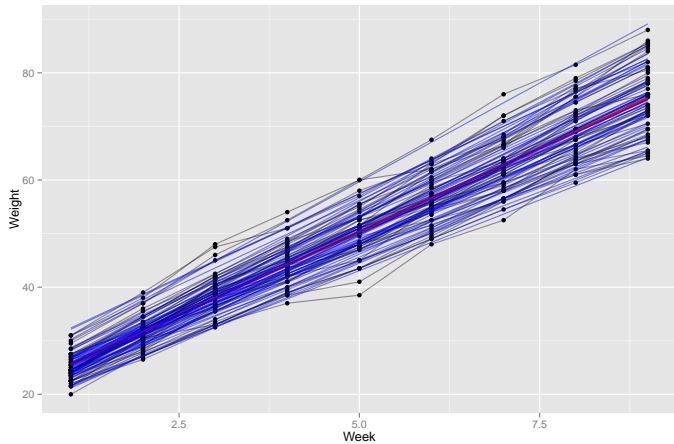
```
> library(lme4)
> ranef.mod = lmer(weight ~ (1 | id.num) + num.weeks, data = pig.weights)
> summary(ranef.mod)
Linear mixed model fit by REML
Formula: weight ~ (1 | id.num) + num.weeks
Data: pig.weights
   AIC   BIC logLik deviance REMLdev
2042 2058 -1017   2030    2034
Random effects:
  Groups   Name      Variance Std.Dev.
id.num    (Intercept) 15.1418  3.8913
Residual                4.3947  2.0964
Number of obs: 432, groups: id.num, 48

Fixed effects:
              Estimate Std. Error t value
(Intercept) 19.35561    0.60311   32.09
num.weeks   6.20990    0.03906  158.97

> (15.1418) / (15.1418 + 4.3947)
[1] 0.7750518
```

# Pig data

Random slope model fit for pig data





# Pig data

## Random slope model code

```
> ranef.mod = lmer(weight ~ (1 + num.weeks | id.num) + num.weeks, data = pig.weights)
```

```
> summary(ranef.mod)
```

Linear mixed model fit by REML

Formula: weight ~ (1 + num.weeks | id.num) + num.weeks

Data: pig.weights

AIC	BIC	logLik	deviance	REMLdev
-----	-----	--------	----------	---------

1753	1777	-870.4	1738	1741
------	------	--------	------	------

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
id.num	(Intercept)	6.9865	2.64319	
	num.weeks	0.3800	0.61644	-0.063
Residual		1.5968	1.26366	

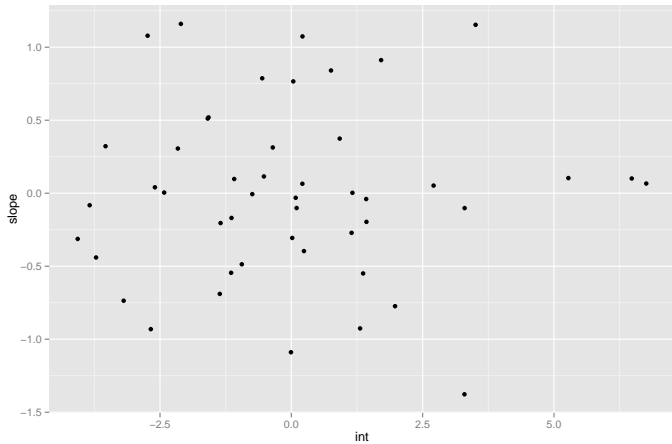
Number of obs: 432, groups: id.num, 48

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	19.35561	0.40387	47.93
num.weeks	6.20990	0.09204	67.47

# Pig data

Random intercept against random slope



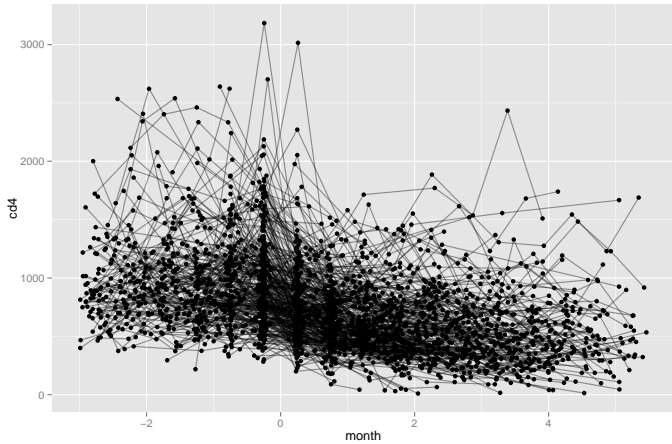
## Quick questions

What data would lead to the random intercept and random slope having high positive correlation? High negative correlation?

## Pig data summary

- Overall, the random slope model provides a pretty good fit
- Lowest AIC of all models considered (linear model not shown, but trust me)
- Visual inspection of data indicates a good fit
- Easy to interpret

# CD4 data



# CD4 data

## SLR model code

```
> lin.mod = lm(cd4 ~ month, data = cd4)
> summary(lin.mod)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	839.398	8.147	103.03	<2e-16 ***
month	-89.027	3.965	-22.45	<2e-16 ***

---

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

Residual standard error: 362.9 on 2369 degrees of freedom

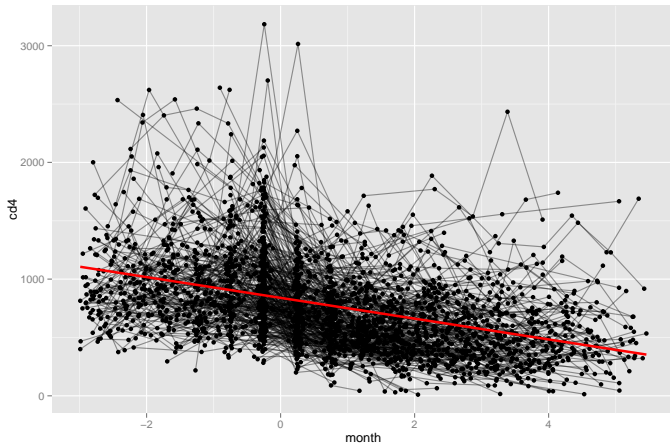
Multiple R-squared: 0.1754, Adjusted R-squared: 0.1751

```
> AIC(lin.mod)
```

```
[1] 34682.64
```

# CD4 data

## SLR model fit for CD4 data



# CD4 data

## B spline model code

```
> bs.mod = lm(cd4 ~ bs(month, 5), data = cd4)
> summary(bs.mod)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	898.19	57.65	15.580	< 2e-16	***
bs(month, 5)1	181.75	101.47	1.791	0.07340	.
bs(month, 5)2	154.21	57.27	2.693	0.00713	**
bs(month, 5)3	-544.31	84.28	-6.459	1.28e-10	***
bs(month, 5)4	-230.25	80.16	-2.873	0.00411	**
bs(month, 5)5	-400.92	98.69	-4.063	5.01e-05	***

---

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

Residual standard error: 356.2 on 2365 degrees of freedom

Multiple R-squared: 0.2068, Adjusted R-squared: 0.2051

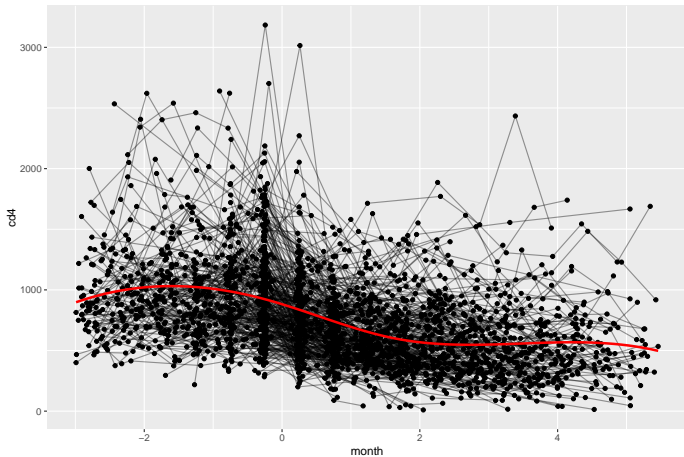
```
> AIC(bs.mod)
```

```
[1] 34598.69
```



# CD4 data

## Polynomial model fit for CD4 data



# CD4 data

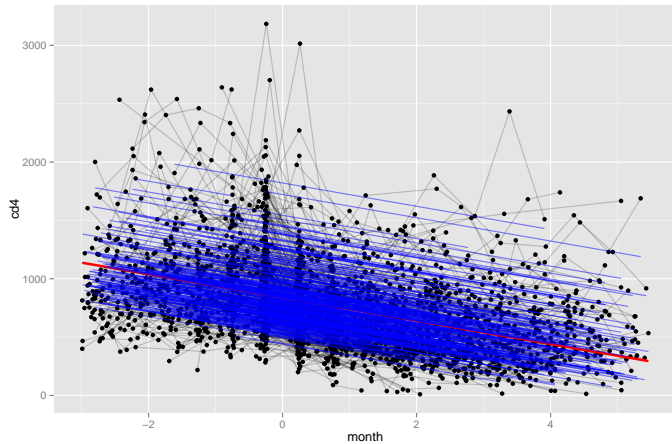
## Random intercept model code

```
> ranint.mod = lmer(cd4 ~ (1 | ID) + month, data = cd4)
> summary(ranint.mod)
Random effects:
  Groups   Name      Variance Std.Dev.
  ID      (Intercept) 64982    254.92
  Residual                    66532    257.94
Number of obs: 2371, groups: ID, 364

Fixed effects:
              Estimate Std. Error t value
(Intercept)  838.925    14.724   56.98
month        -99.703     3.449  -28.91
> AIC(ranint.mod)
[1] 33755.67
```

# CD4 data

## Random intercept fit for CD4 data



# CD4 data

## Random intercept, slope + B spline model code

```
> ranbs.mod = lmer(cd4 ~ (1 + month | ID) + bs(month, 5), data = cd4)
> summary(ranbs.mod)
```

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
ID	(Intercept)	71333	267.08	
	month	4851	69.65	-0.43
Residual		50484	224.69	

Number of obs: 2371, groups: ID, 364

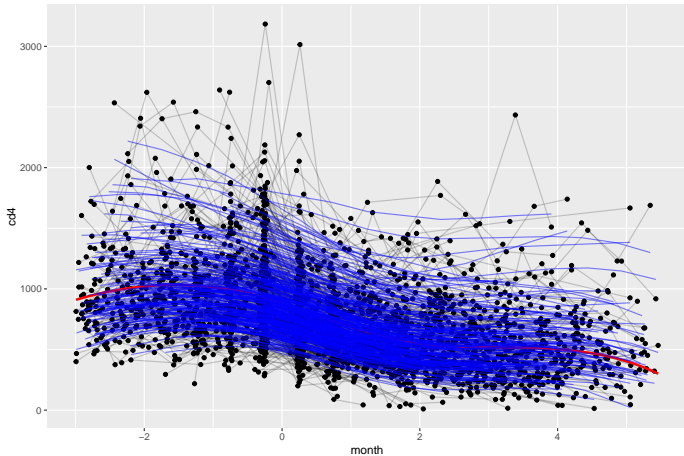
Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	910.11	50.09	18.169
bs(month, 5)1	157.54	73.43	2.145
bs(month, 5)2	164.87	47.42	3.477
bs(month, 5)3	-588.42	64.43	-9.133
bs(month, 5)4	-264.07	65.57	-4.027
bs(month, 5)5	-609.54	82.14	-7.421

```
> AIC(ranbs.mod)
[1] 33428.49
```

# CD4 data

Random intercept, slope + B spline fit for CD4 data



# CD4 data summary

Which model do you prefer?

# Today's big ideas

- Random slope models
  - Pig data analysis
  - CD4 data analysis
-