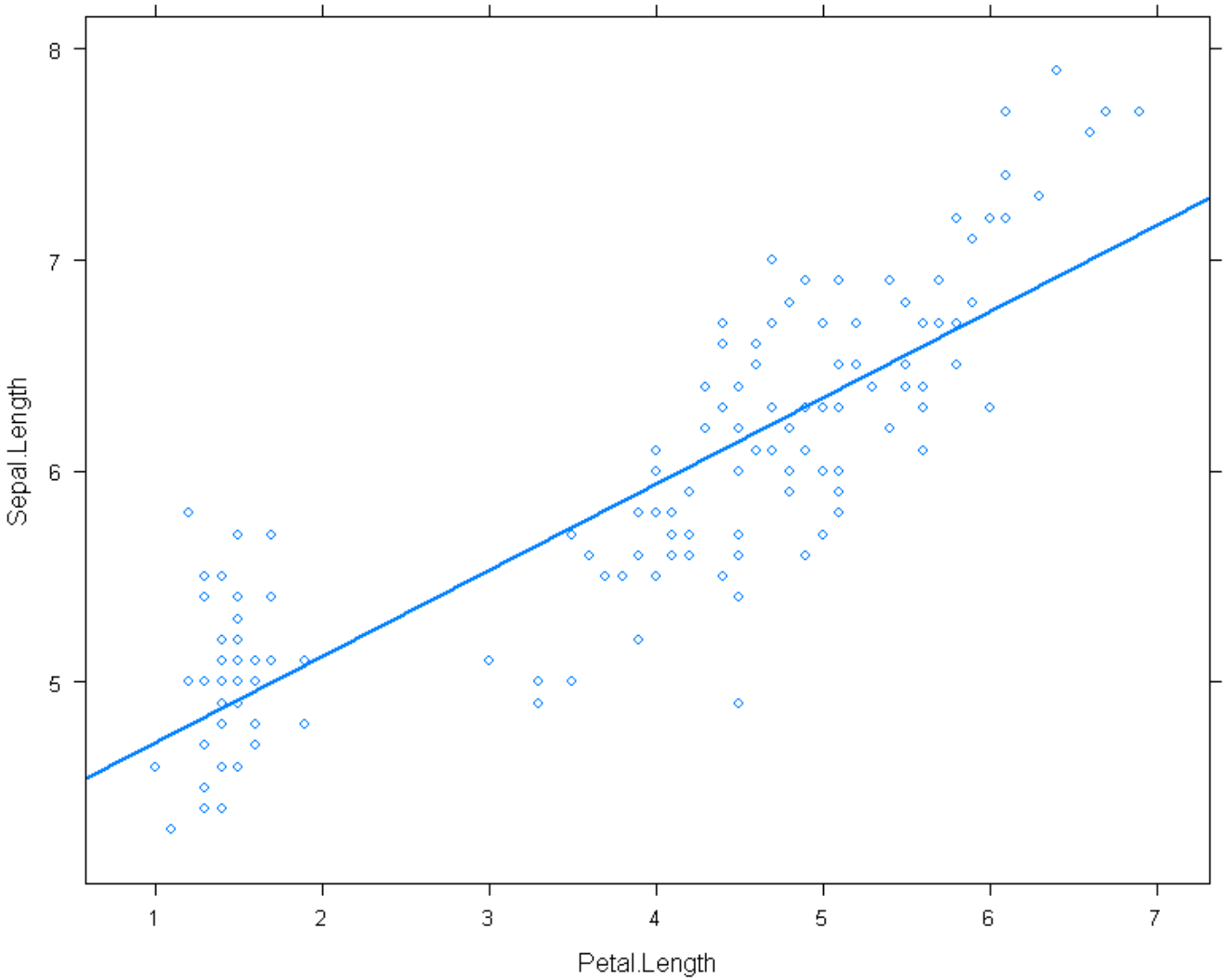
Regression and Hierarchical (Mixed) Models

Bob Farmer

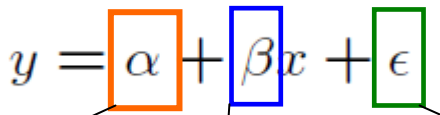
farmerb@dal.ca

http://leonardlab.biology.dal.ca


Sepal.Length
Petal.Length
8
7
6
5
1
2
3
4
5
6
7

Regression – the non-matrix version

- Estimate y in terms of parameters β using dataset X



Intercept: Value of y if x is zero

Slope(s): The increase in *y* (above α) for every unit increase in *x*

Model error (residuals): The difference between each actual and predicted value (cannot be estimated)

$$Residuals = y_i - \widehat{y_i}$$

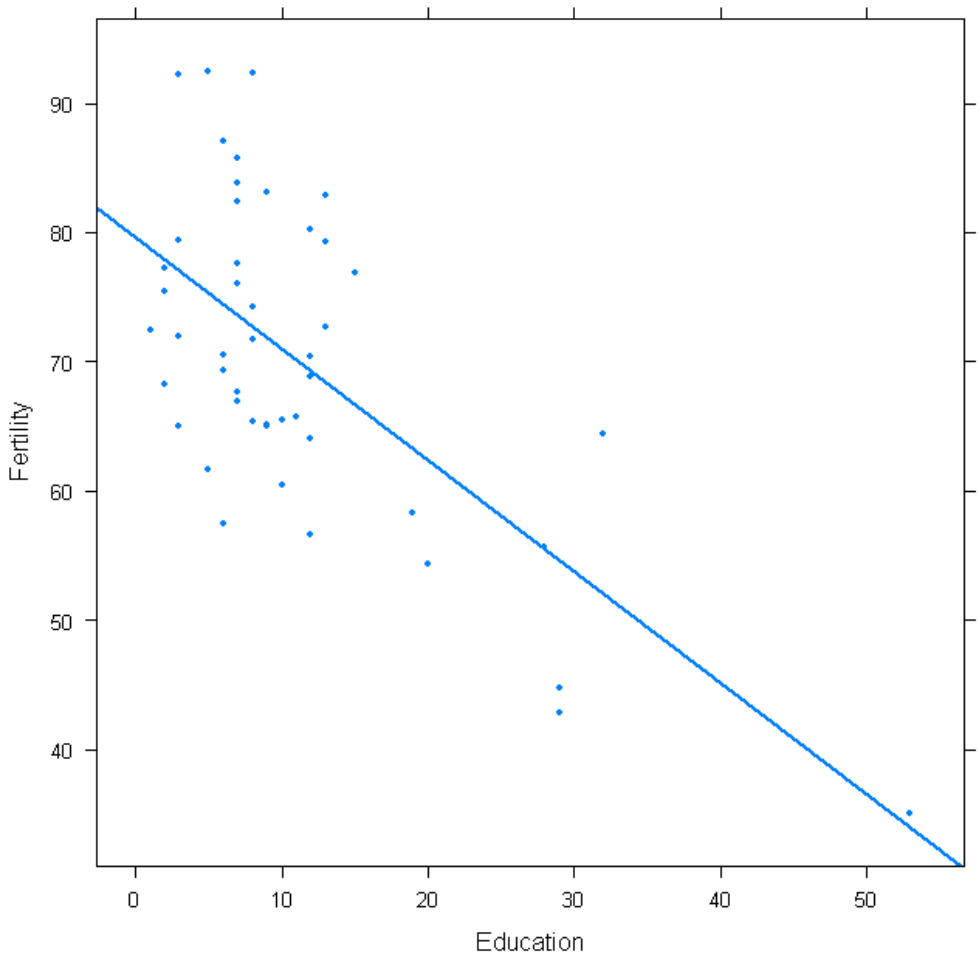
```
> swiss[1:6,c(1,4)]
             Fertility Education
Courtelary        80.2        12
Delemont          83.1         9
Franches-Mnt      92.5         5
Moutier           85.8         7
Neuveville        76.9        15
Porrentruy        76.1         7
```

```
data(swiss)
m0<-lm(Fertility ~ Education, swiss)
summary(m0)
...
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  79.6101     2.1041  37.836  < 2e-16 ***
Education    -0.8624     0.1448  -5.954 3.66e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 9.446 on 45 degrees of freedom
```

$$y = \alpha + \beta x + \epsilon$$

$$y = 79.6101 + (-0.8624)x \pm 9.446$$

```
library(lattice)
xyplot(Fertility ~ Education, swiss,
type = c("p", "r"), lwd = 2, pch = 20
)
```



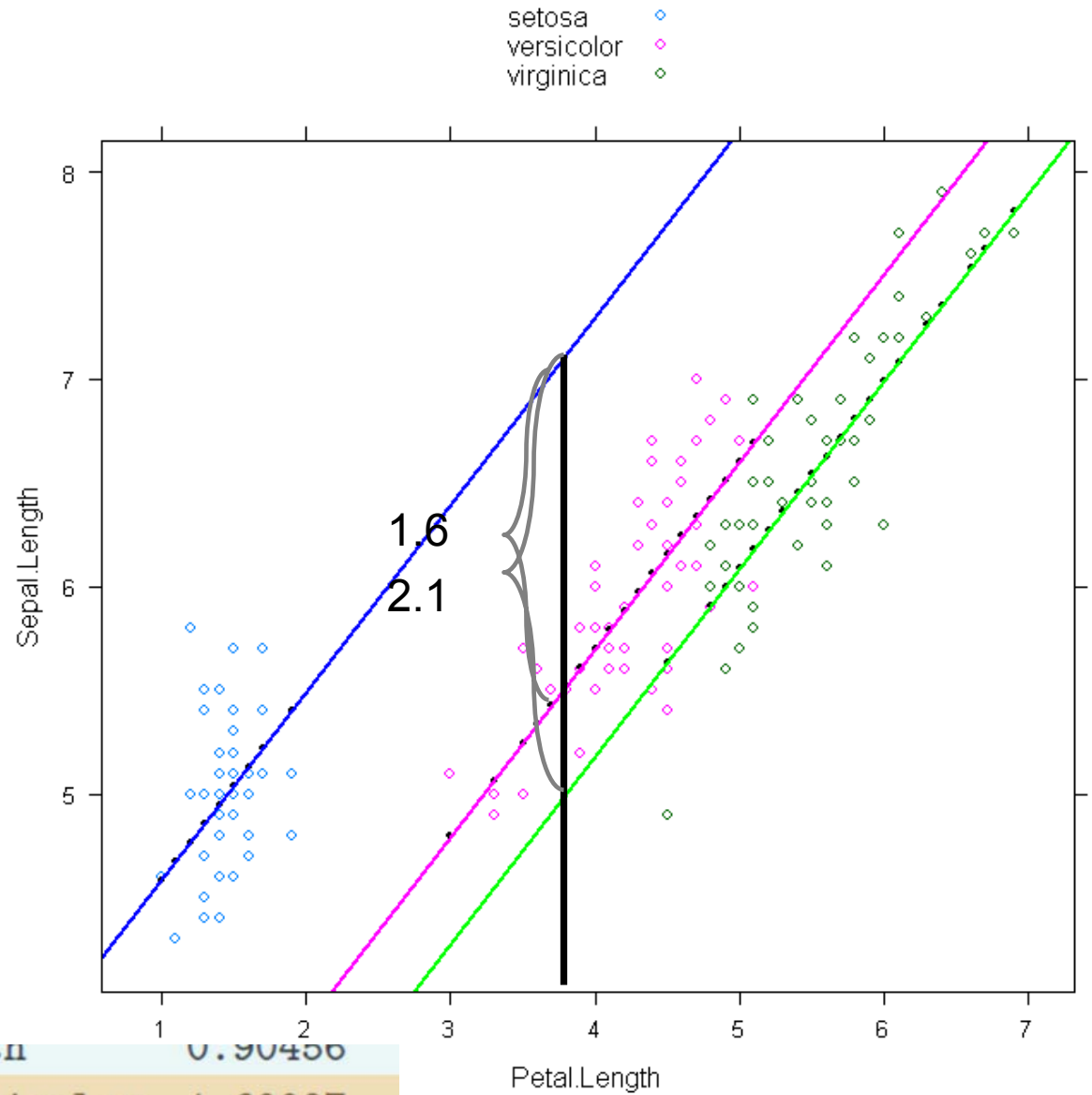
Groups

- What if we add additional, *categorical* predictors (e.g. a known grouping structure)?

```
> iris[c(1:10, 50:60),c(1,3,5)]
   Sepal.Length Petal.Length    Species
1           5.1          1.4     setosa
2           4.9          1.4     setosa
3           4.7          1.3     setosa
4           4.6          1.5     setosa
5           5.0          1.4     setosa
6           5.4          1.7     setosa
7           4.6          1.4     setosa
8           5.0          1.5     setosa
9           4.4          1.4     setosa
10          4.9          1.5     setosa
50          5.0          1.4     setosa
51          7.0          4.7 versicolor
52          6.4          4.5 versicolor
53          6.9          4.9 versicolor
54          5.5          4.0 versicolor
55          6.5          4.6 versicolor
56          5.7          4.5 versicolor
57          6.3          4.7 versicolor
58          4.9          3.3 versicolor
59          6.6          4.6 versicolor
60          5.2          3.9 versicolor
```

Blocking by group (independent *intercepts*)

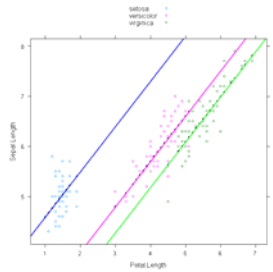
```
m3<-lm(Sepal.Length ~ Petal.Length + Species, iris)
summary(m3)
...
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.68353    0.10610  34.719  < 2e-16 ***
Petal.Length       0.90456    0.06479  13.962  < 2e-16 ***
Speciesversicolor -1.60097    0.19347  -8.275 7.37e-14 ***
Speciesvirginica  -2.11767    0.27346  -7.744 1.48e-12 ***
```




```
Petal.Length      0.90456
Speciesversicolor -1.60097
Speciesvirginica  -2.11767
```

$$y_i = \beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2} + \beta_3 x_{i_3} + \epsilon$$

```
m3<-lm(Sepal.Length ~ Petal.Length + Species, iris)
summary(m3)
...
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.68353  β0 0.10610  34.719  < 2e-16 ***
Petal.Length       0.90456  β1 0.06479  13.962  < 2e-16 ***
Speciesversicolor -1.60097  β2 0.19347  -8.275 7.37e-14 ***
Speciesvirginica  -2.11767  β3 0.27346  -7.744 1.48e-12 ***
```



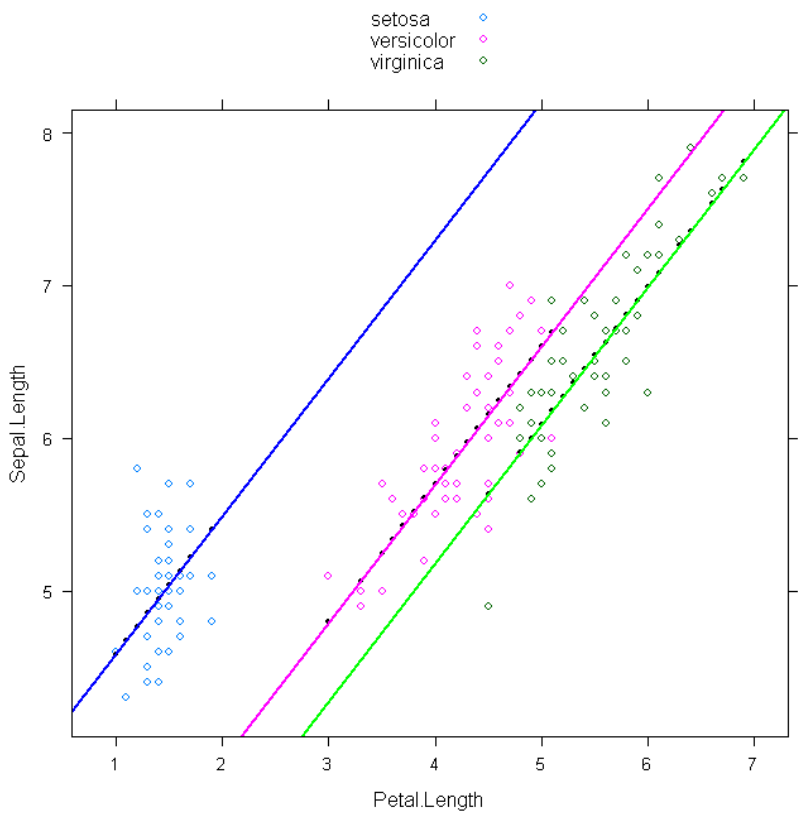
Sepal Length = (Intercept) β_0

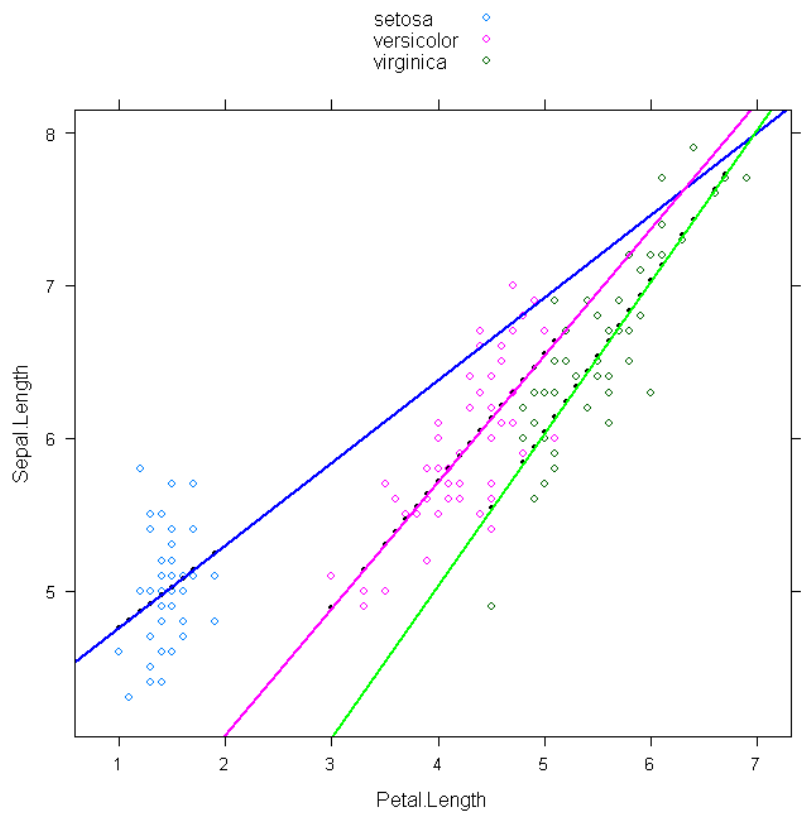
+ (Petal Length Modifier) x (Petal Length) β_1

+ ("versicolor-ness" Modifier) x (versicolor (1 or 0)?) β_2

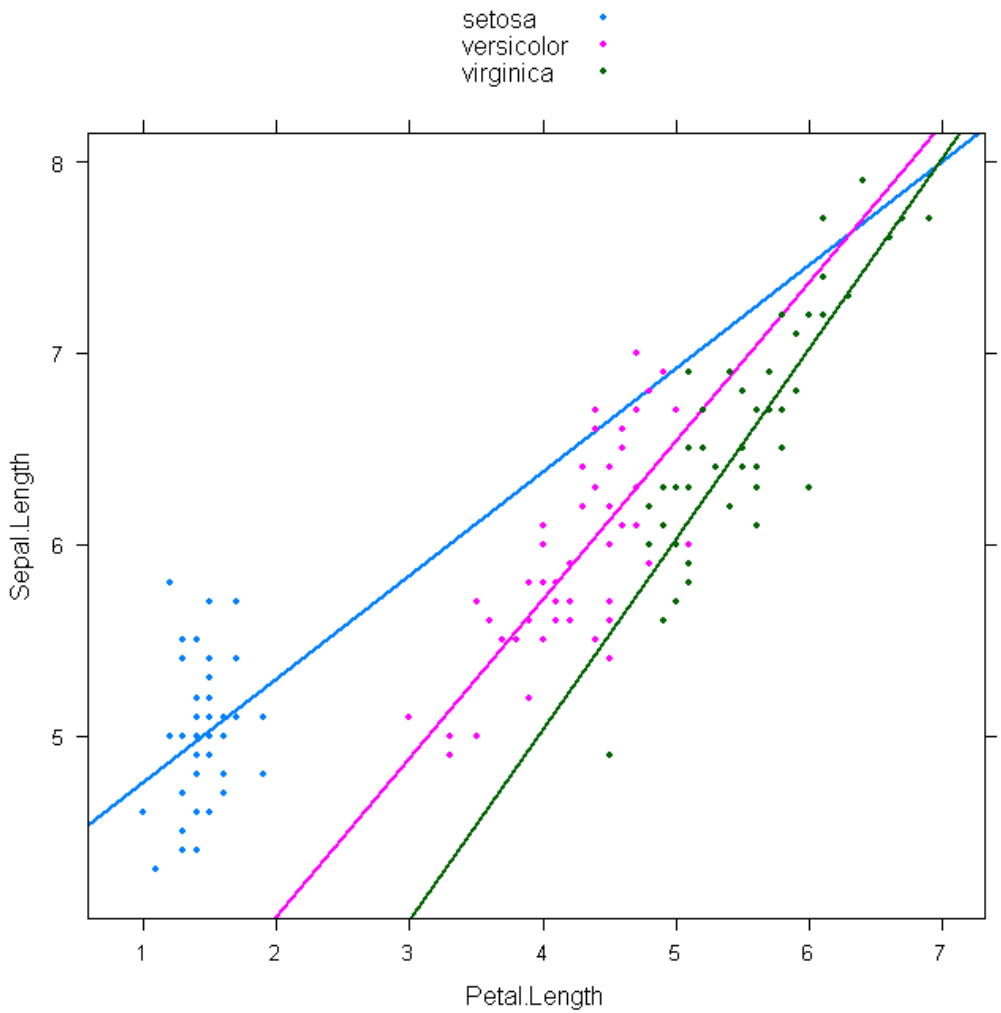
+ ("virginica-ness" Modifier) x (virginica (1 or 0)?) β_3

+ Error

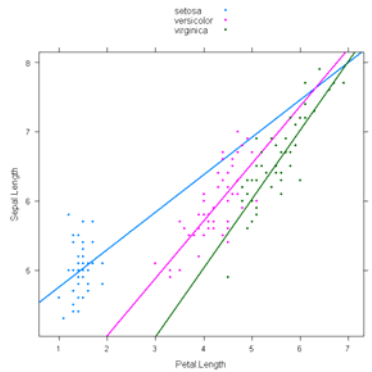

setosa
versicolor
virginica
Sepal.Length
Petal.Length
8
7
6
5
1
2
3
4
5
6
7


setosa
versicolor
virginica
Sepal.Length
Petal.Length
8
7
6
5
1
2
3
4
5
6
7

Three separate regressions (*interaction;* independent *slopes* and *intercepts*)



Interaction (independent *slopes* and *intercepts*)




```
data(iris)
m2<-lm(Sepal.Length ~ Petal.Length*Species, iris)
summary(m2)
...
Coefficients:
```

	Estimate		Std. Error	Pr(>|t|)
(Intercept)	4.21	β_0	0.41	<2e-16 ***
Petal.Length	0.54	β_1	0.28	0.052 .
Speciesversicolor	-1.81	β_2	0.60	0.003 **
Speciesvirginica	-3.15	β_3	0.63	2e-06 ***
Petal.Length:Speciesversicolor	0.29	β_4	0.30	0.334
Petal.Length:Speciesvirginica	0.45	β_5	0.29	0.120

$$y_i = \beta_0 + (\beta_1 + \beta_4 x_{i_2} + \beta_5 x_{i_3}) x_{i_1} + \beta_2 x_{i_2} + \beta_3 x_{i_3} + \epsilon$$

Independent Slopes or Intercepts?

- More complicated models:
  - Need a greater sample size
  - Are better at prediction within their native systems
  - Are not always necessary

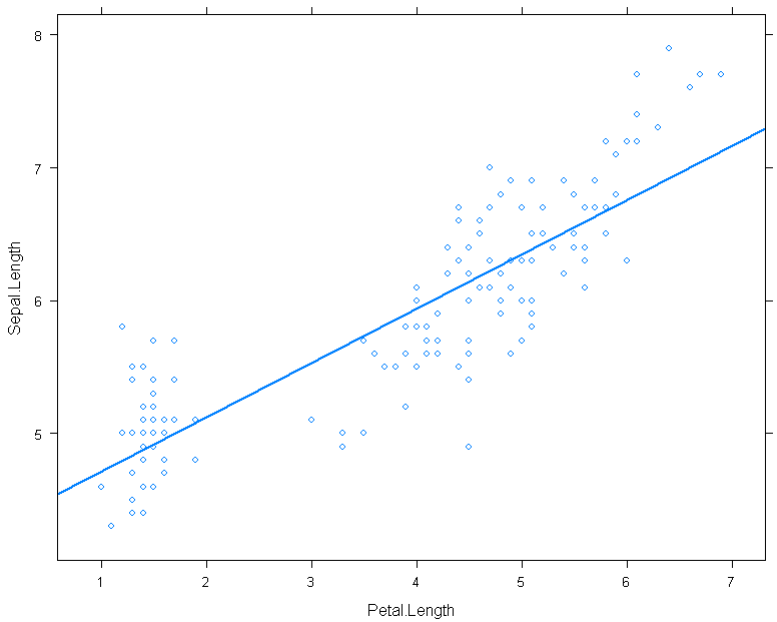

VOGUE
3,50 €
NOVIEMBRE
ESPAÑA 2007
CONTENIDO ADULTO
137
MOMENTOS DE PLACER
EL VESTIDO TUBO
LA FALDA DE TWEED
EL PANTALÓN ANCHO
LA CHAQUETA MASCULINA...
TE HACEN SEXY
SIEMPRE PERFECTA
...Y TAN FÁCIL!
10 BÁSICOS
¡SIN REMORDIMIENTOS!
LA PIEL QUE SE LLEVA
BOLERO DE ZORRO + VESTIDO
CHAQUETÓN DE PELO LARGO + MINI
CHALECO DE VISÓN + PITILLO
Y EL MEJOR PEELING
PROBAMOS EL ÚLTIMO LÁSER MILAGRO
MUJERES EXTREMAS
ANA CAMAZO
SU VIDA EN ÁFRICA
CLAUDIA SCHIFFER
DESNUDA A LOS 37

Some Philosophy…

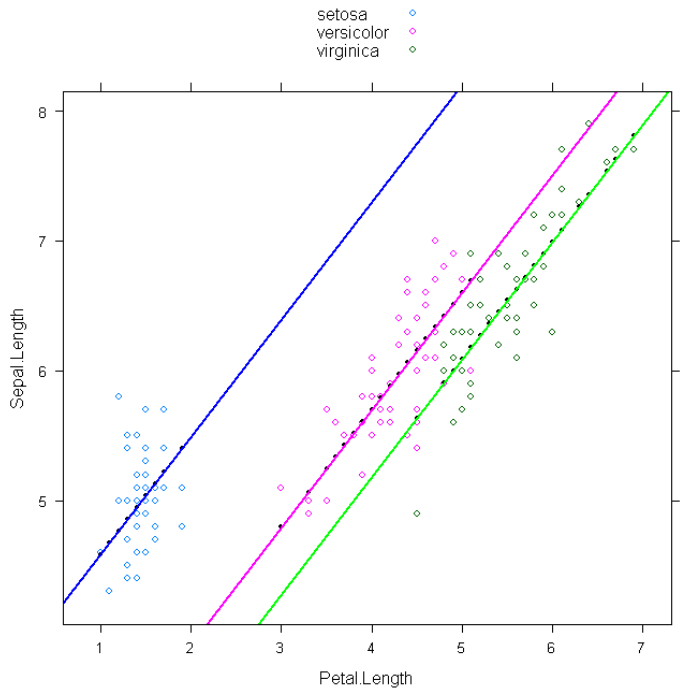
- Suppose we are aware of a grouping structure (i.e. “Species”), but:
  - i) we don’t care about the identity and effects of each individual group; and/or
  - ii) there are more potential groups out there than we can measure
    - e.g. “Mother ID”, “Growth Plot”, “Region”
- Specify group effects “in the background”, focusing instead on the “main” effects
  - e.g. the change in Sepal Length with Petal Length

The Hierarchical "Compromise"

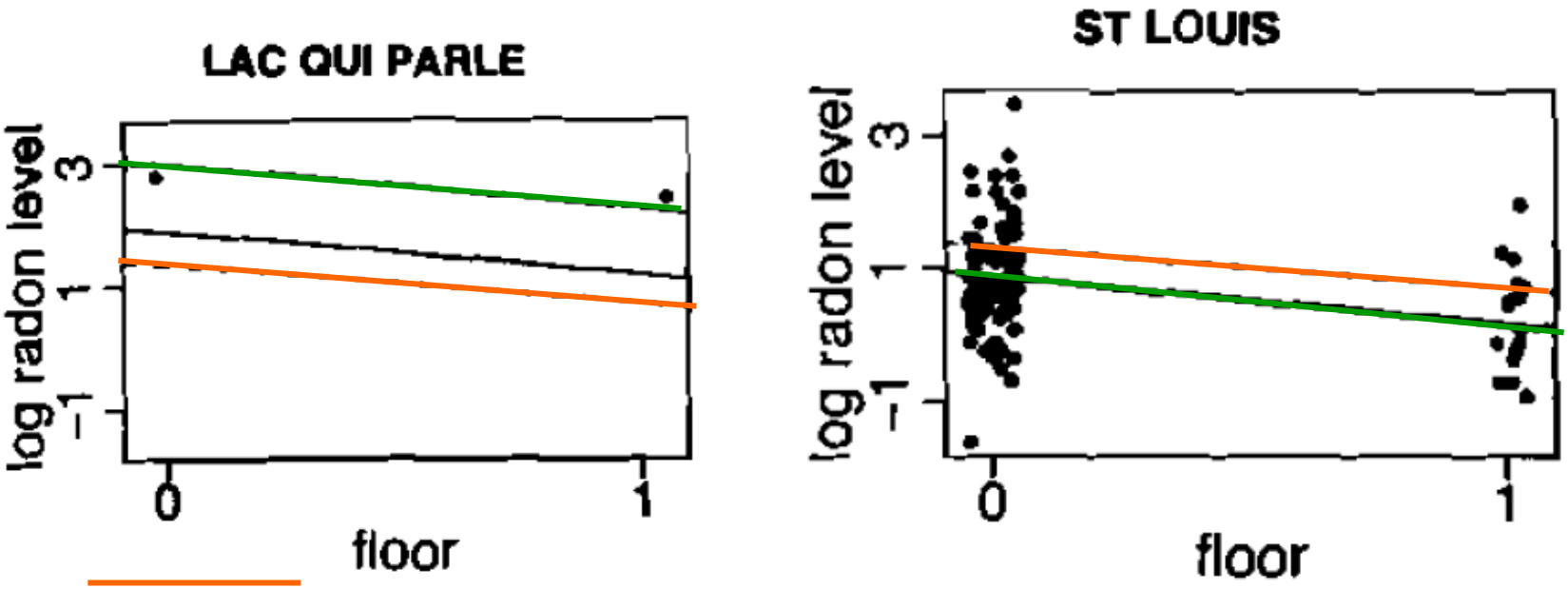
- Imagine a model that fits between these two modeling (*intercept*) extremes:



"Global" (One intercept)



"Individual" (several intercepts)



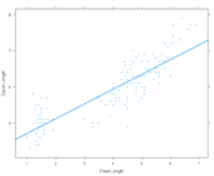
"Global" (one intercept) model

"Individual" (several intercepts) model

Hierarchical model fit

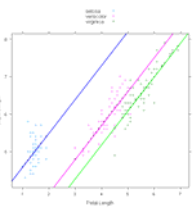
Choose an intercept for each group, but dampen the effect (and be sensitive to group sample size)

Gelman and Hill 2007



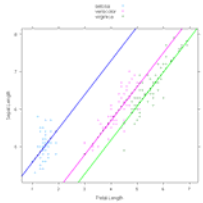
"Global" (one intercept) model

"Individual" (several intercepts) model



Hierarchical model fit

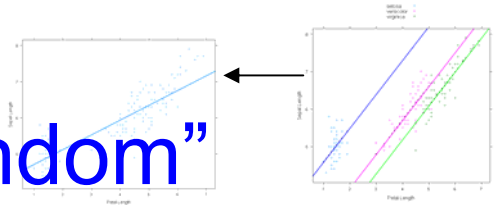
- The dampened group intercepts come from a distribution of possible values



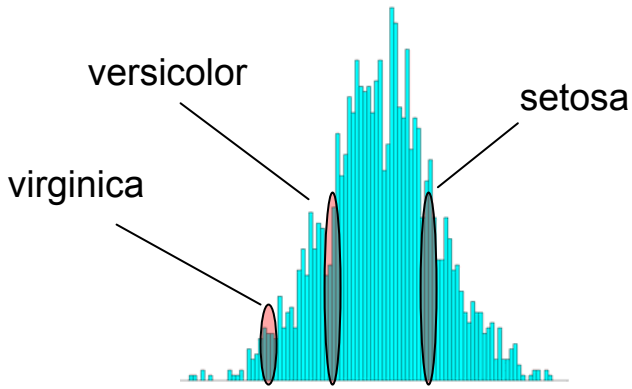
“Fixed” effects

versicolor	-1.6
virginica	-2.1

Can be *any* value that best fits the results



“Random” (hierarchical model) effects



Values from a defined distribution (mean is usually the “global”-model intercept)

Hierarchical (Random Effects) Model Formulation

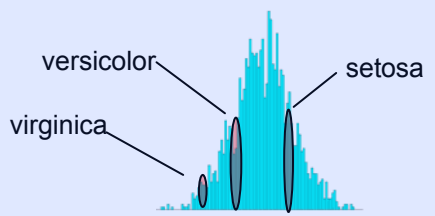
- Hierarchical ("random effects", "multilevel") models

$$y = \beta_0 + \beta_1 x + \epsilon$$

now allow intercept to vary within groups j:

$$y = \beta_{0j} + \beta_1 x + \epsilon$$ Hierarchical ("random intercept")

where $\beta_{0j} \sim N(\mu_{\beta_0}, \sigma^2_{\beta_0})$

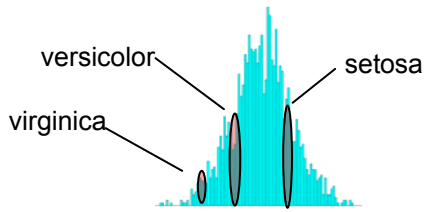


with error η_j

Hierarchical Models: The Hierarchy

- The group-level predictors (β_j) are derived from a (restricted) model

$$\beta_{0j} \sim N(\mu_{\beta_0}, \sigma^2_{\beta_0})$$



implies

$$\beta_{0j} = a_0 + b_0\mu_j + \eta_{j1}$$

within

$$y = \beta_{0j} + \beta_1 x + \epsilon$$

Random Slopes

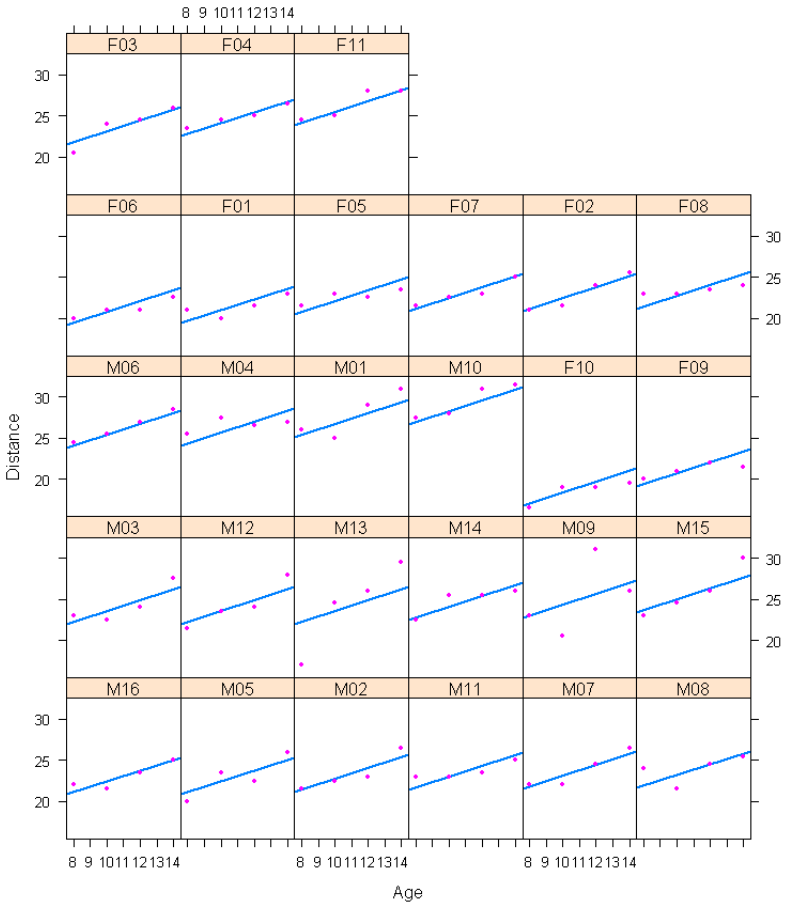
- For a more complicated (and specific) model, slopes can also vary hierarchically:

$$y = \beta_{0j} + \beta_{1j}x + \epsilon$$

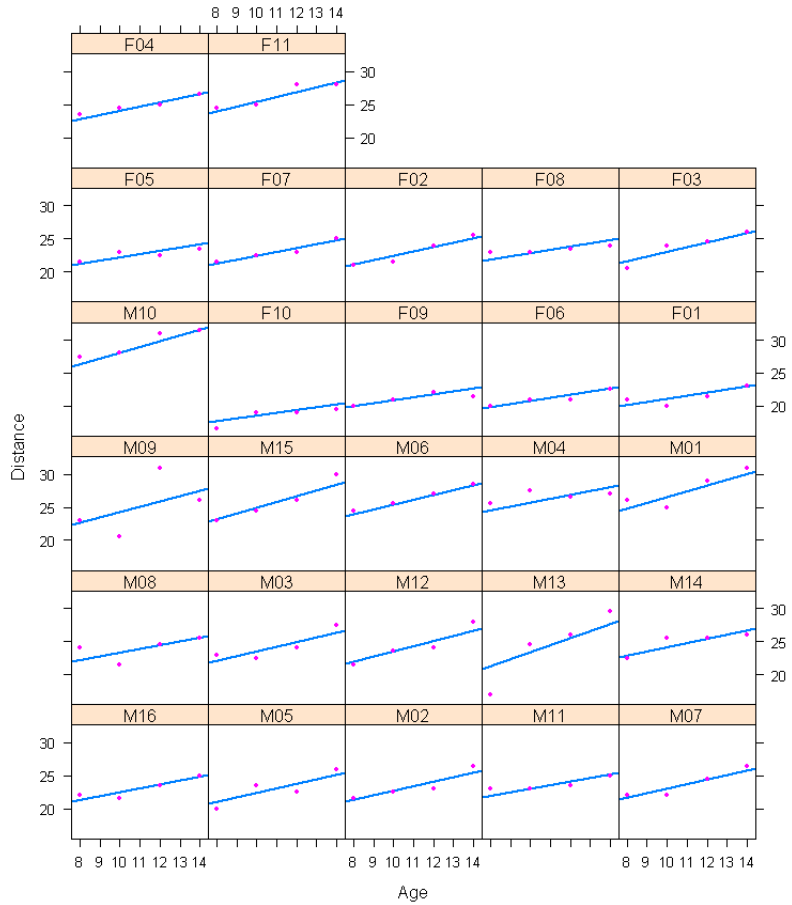
where $\beta_{0j} \sim N(\mu_{\beta_0}, \sigma^2_{\beta_0})$

and $\beta_{1j} \sim N(\mu_{\beta_1}, \sigma^2_{\beta_1})$

with errors $\eta_j^{\beta_1}$ and $\eta_j^{\beta_2}$



Random Intercepts

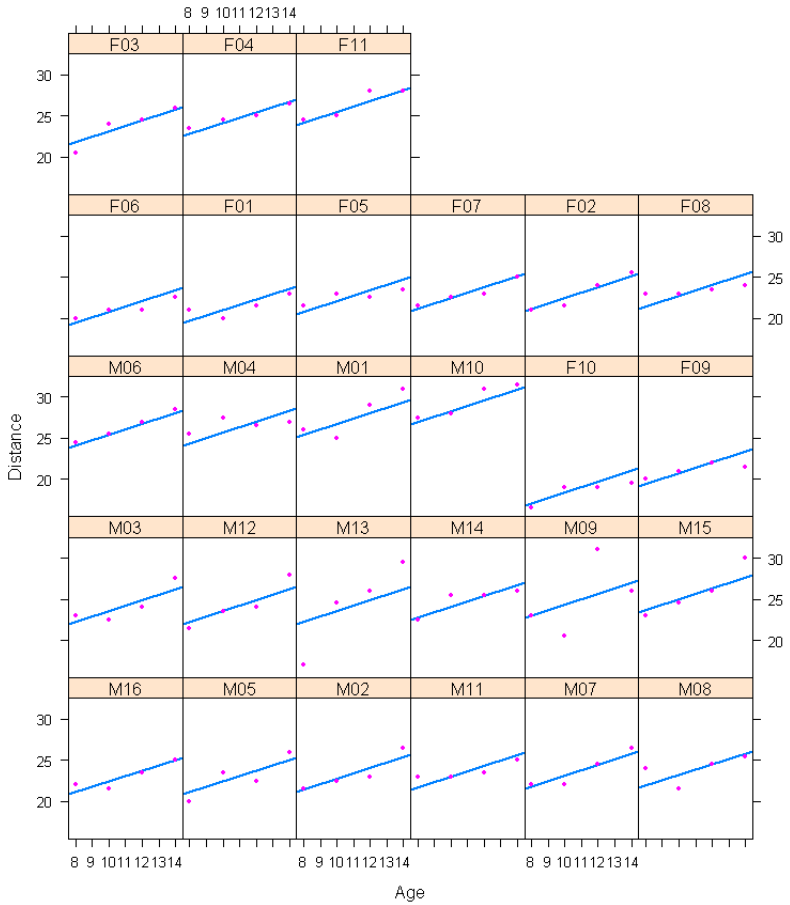


Random Slopes and Intercepts

The values of the random intercepts and slopes are not as variable as in a fixed-effects model

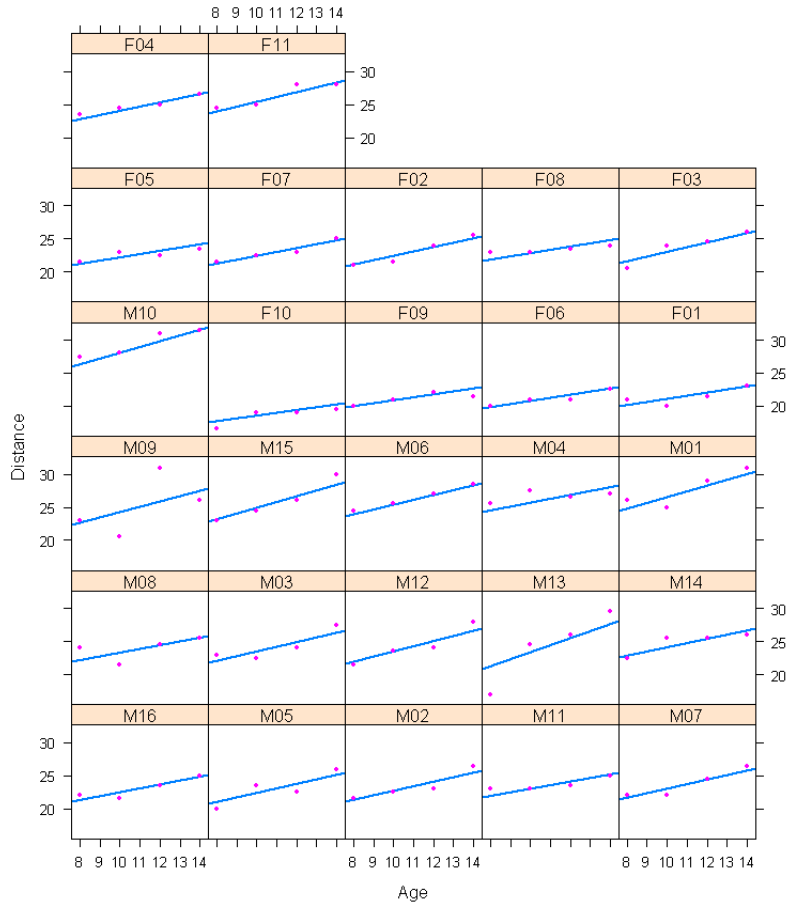
R and lme()

- lme() in package *nlme*
- Extra argument is “random”
  - i.e. random = ~ 1 | groups
  - Variables to the left of the “|” are slopes; to the right are intercepts
    - E.g. Random Intercepts:
      - random = ~1 | groups
    - Random Slopes:
      - random = ~ someVar | groups



Random Intercepts

```
m5<-lme(distance ~ age, random = ~ 1 | Subject, Orthodont)
```



Random Slopes and Intercepts

```
data(Orthodont); library(nlme)
m6<-lme(distance ~ age, random = ~ age | Subject, Orthodont)
```

Link functions and mixed models

- Use glmmPQL (package MASS) or lmer() (package lme4)
  - Analogous to glm() vs. lm()
- I use modified Poisson links for my population data, which has a natural multiplicative structure to it

Simple ("Fixed-effects") Linear Model Output

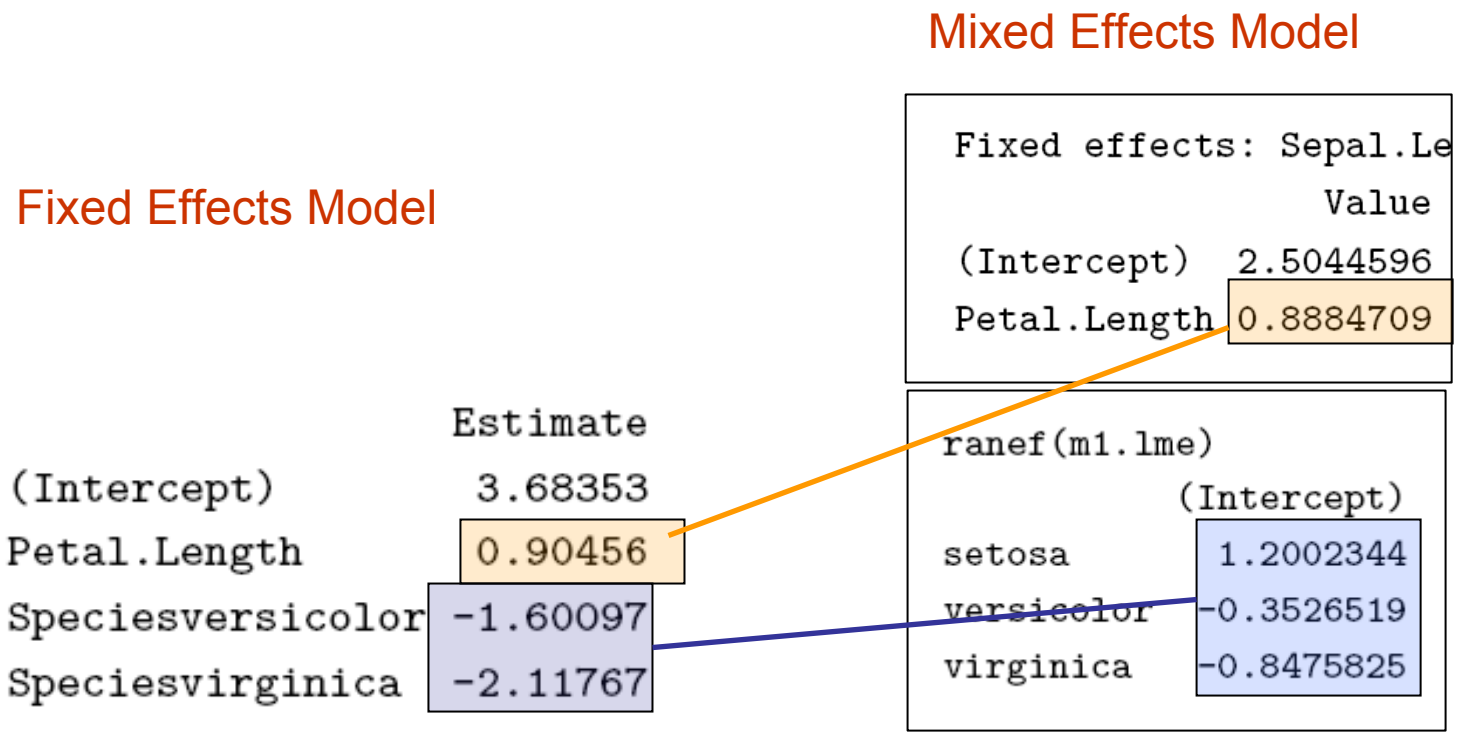
```
m1<-lm(Sepal.Length ~ Petal.Length + Species, iris)
summary(m1)
...
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.68353    0.10610  34.719  < 2e-16 ***
Petal.Length        0.90456    0.06479  13.962  < 2e-16 ***
Speciesversicolor -1.60097    0.19347  -8.275 7.37e-14 ***
Speciesvirginica  -2.11767    0.27346  -7.744 1.48e-12 ***
Residual standard error: 0.338 on 146 degrees of freedom
```

Hierarchical Model Output

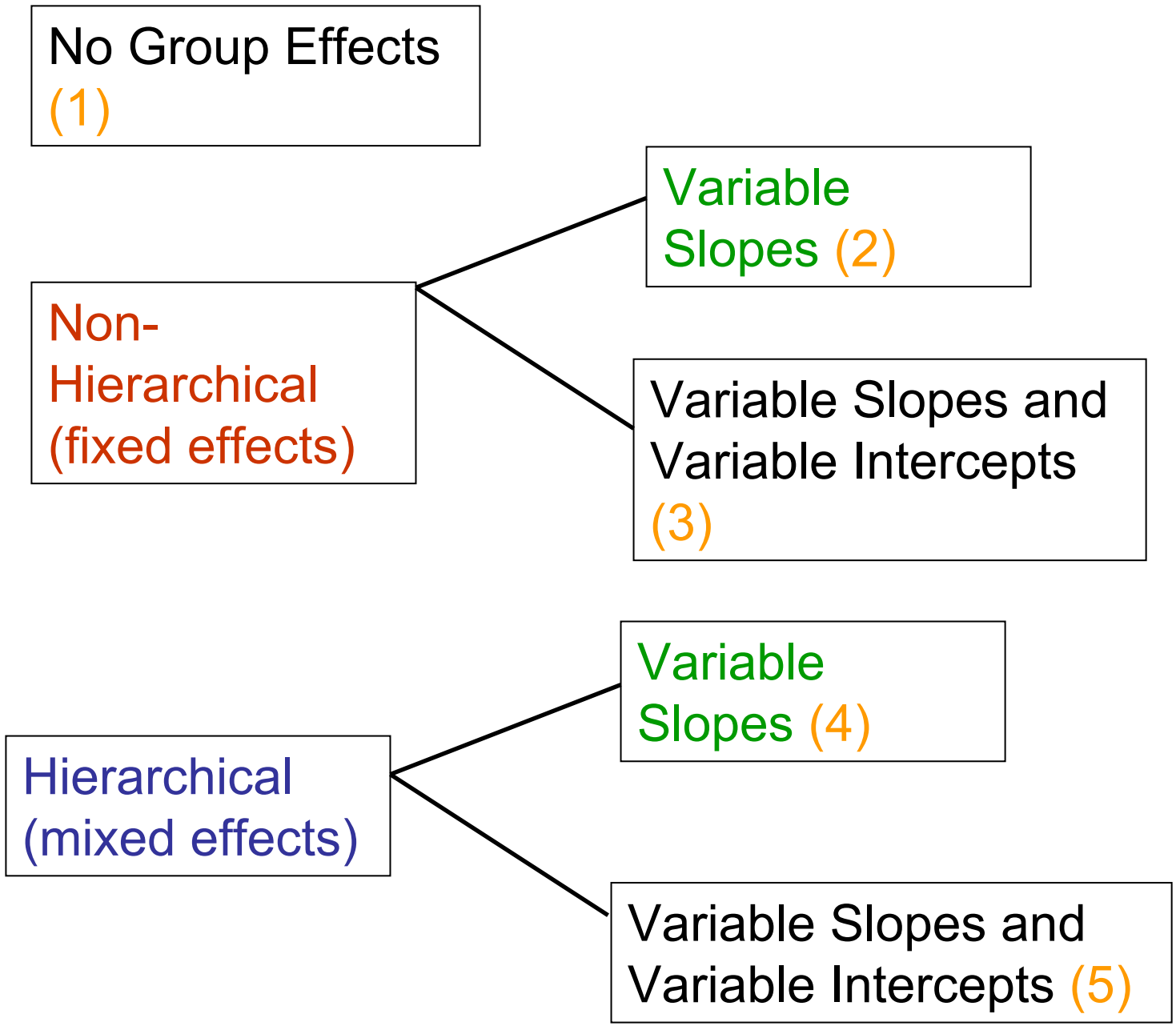
```
m1.lme<-lme(Sepal.Length ~ Petal.Length,
random = ~ 1 | Species, iris)
summary(m1.lme)
...
Random effects:
 Formula: ~1 | Species
        (Intercept)  Residual
StdDev:    1.077804 0.3380567

Fixed effects: Sepal.Length ~ Petal.Length
                  Value Std.Error  DF   t-v
(Intercept)  2.5044596 0.6674261 146  3.75
Petal.Length 0.8884709 0.0637946 146 13.92
```

```
ranef(m1.lme)
           (Intercept)
setosa       1.2002344
versicolor  -0.3526519
virginica   -0.8475825
```


Mixed Effects Model
Fixed effects: Sepal.Le
Value
(Intercept) 2.5044596
Petal.Length 0.8884709
Fixed Effects Model
Estimate
(Intercept) 3.68353
Petal.Length 0.90456
Speciesversicolor -1.60097
Speciesvirginica -2.11767
ranef(m1.lme)
(Intercept)
setosa 1.2002344
versicolor -0.3526519
virginica -0.8475825

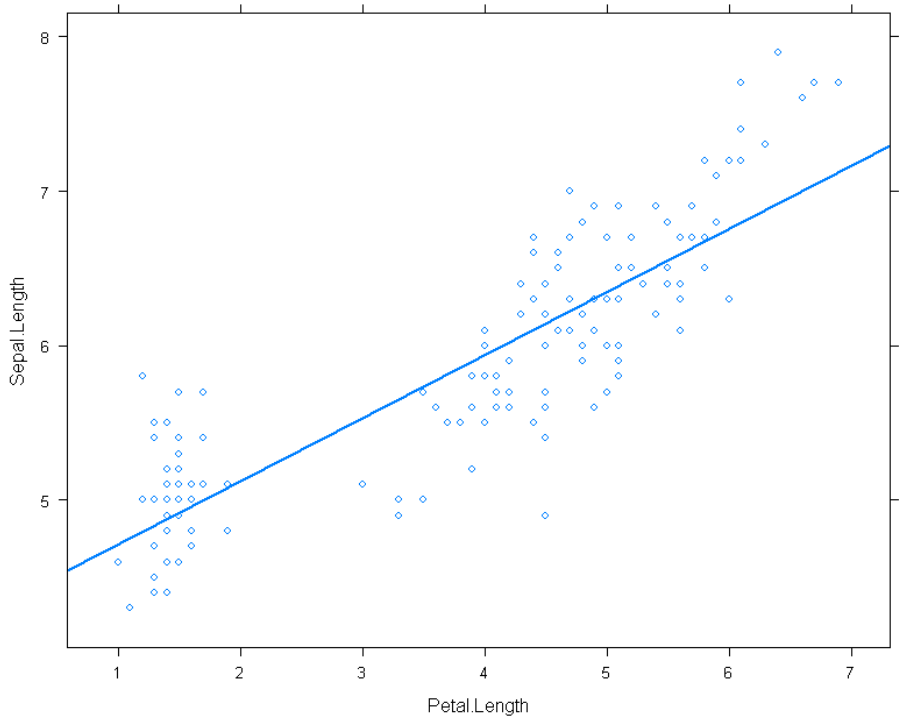
5 ways to approach grouped data


No Group Effects (1)
Non-Hierarchical (fixed effects)
Variable Slopes (2)
Variable Slopes and Variable Intercepts (3)
Hierarchical (mixed effects)
Variable Slopes (4)
Variable Slopes and Variable Intercepts (5)

Iris Data: 5 Different Approaches

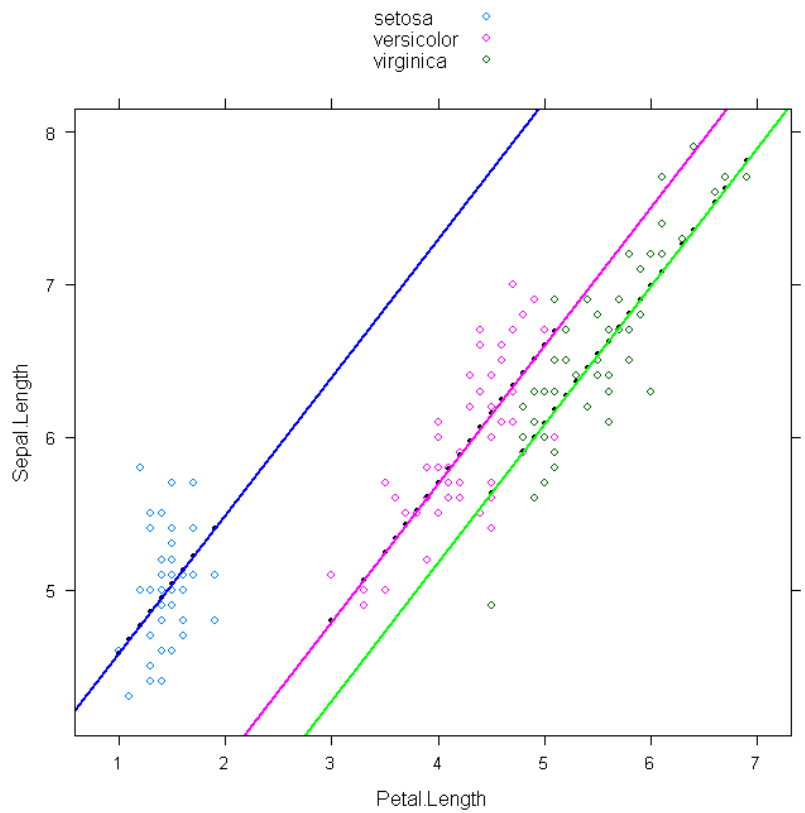
- I want to know the relationship between sepal length and petal length for a set of 3 species

- 1. I don't believe that there are important group-level (species) effects
  - Use global linear model



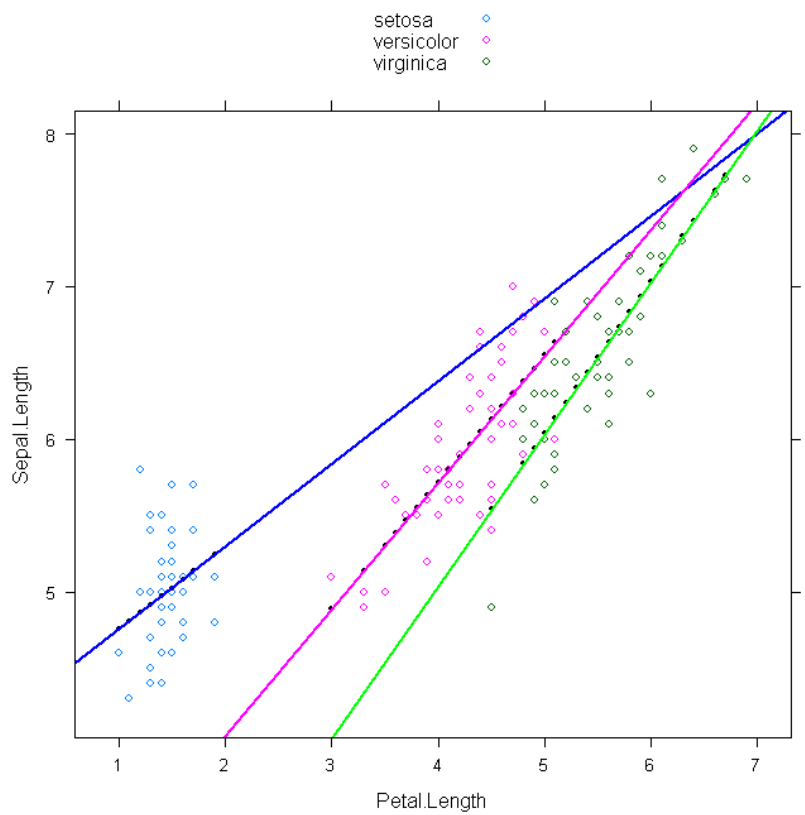

```
mGlob<-lm(Sepal.Length ~ Petal.Length, iris)
```

- 2. I believe that there are important group-level effects
- I care about the values of group-level effects
  - Non-hierarchical model
- I believe that different groups have
  - different means
  - similar rates of change with the independent
    - Variable intercepts (no interaction)



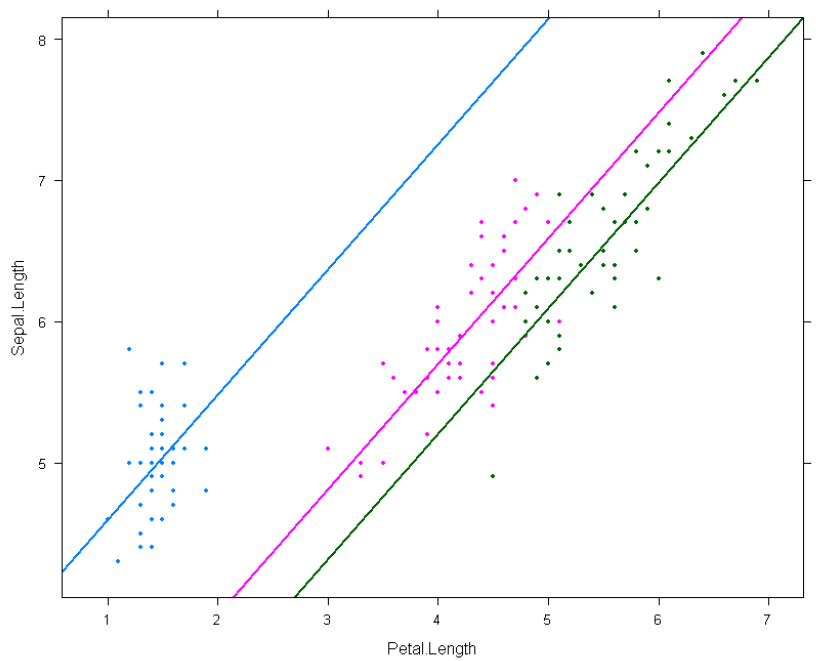

```
mInd_int<-lm(Sepal.Length ~ Petal.Length + Species, iris)
library(nlme)
```

- 3. I believe that there are important group-level effects
- I care about the values of group-level effects
  - Non-hierarchical model
- I believe that different groups have
  - different means
  - different rates of change with the independent
    - Variable slopes (interactions)



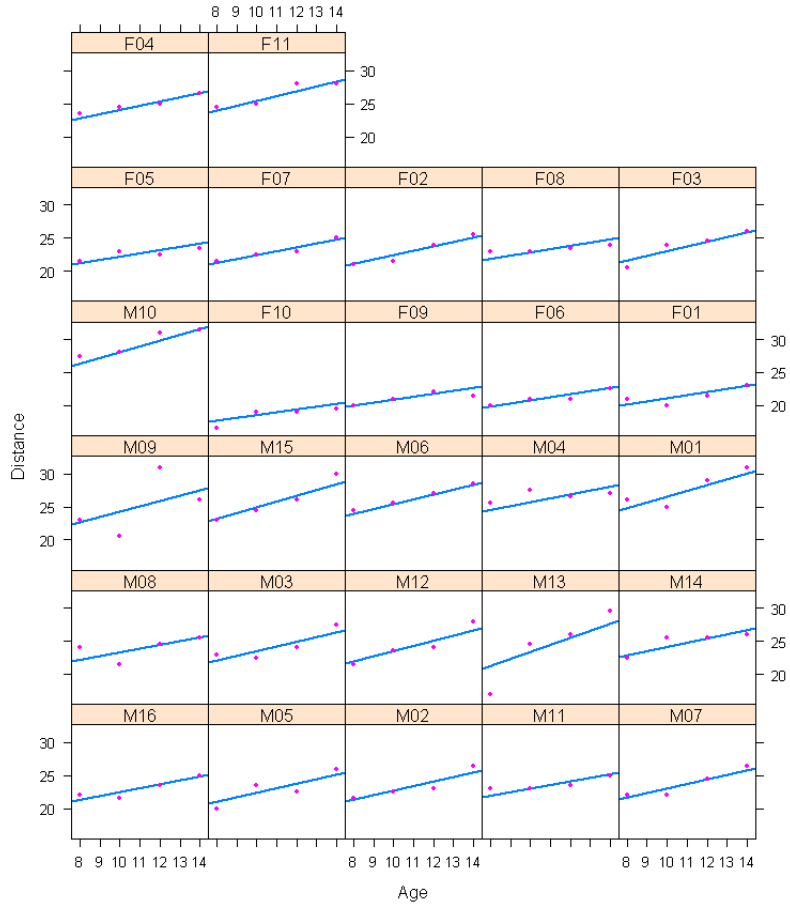

```
mInd_sl<-lm(Sepal.Length ~ Petal.Length + Species
+ Petal.Length:Species, iris)
```

- 4. I believe there are important group-level effects
- I *don't* care about the values of group-level effects
  - Hierarchical model
- I believe that different groups have
  - different means
  - similar rates of change with the independent
    - Random intercepts




```
library(nlme)
mHier_int<-lme(Sepal.Length ~ Petal.Length,
random = ~1|Species, iris)
```

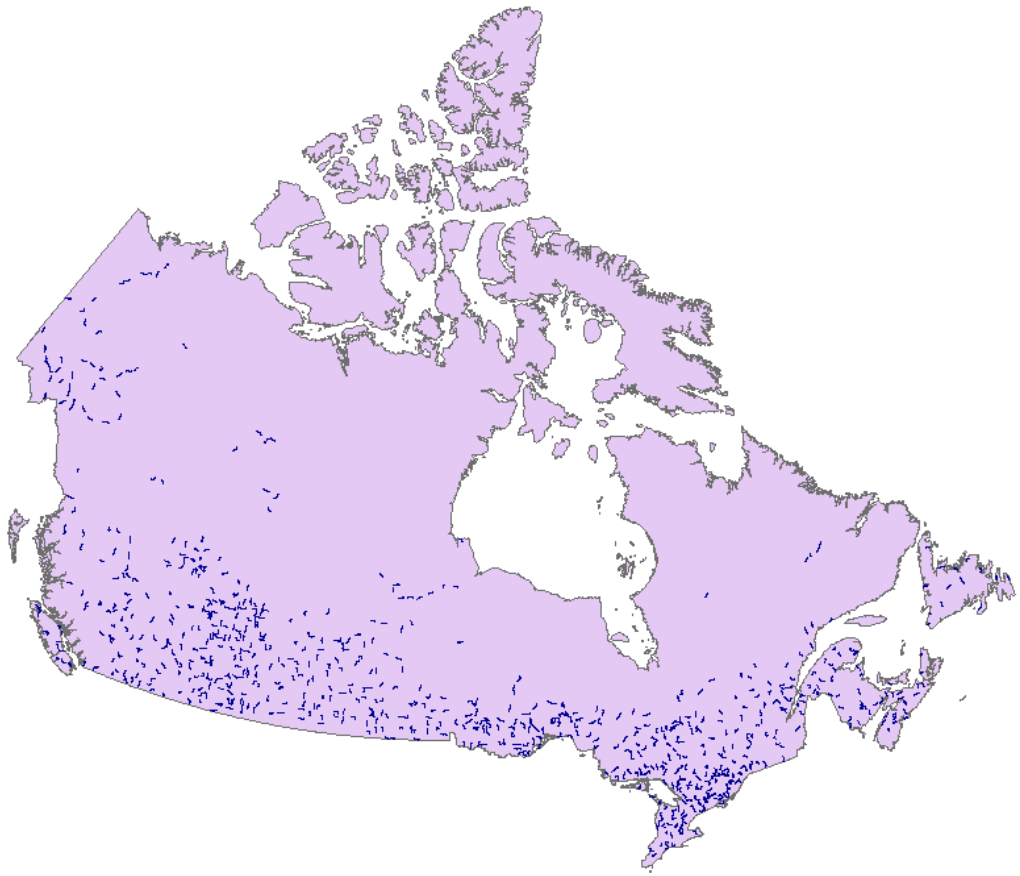
- 5. I believe that there are important group-level effects
- I *don't* care about the values of group-level effects
  - Hierarchical model
- I believe that different groups have
  - different means
  - different rates of change with the independent
    - Random slopes

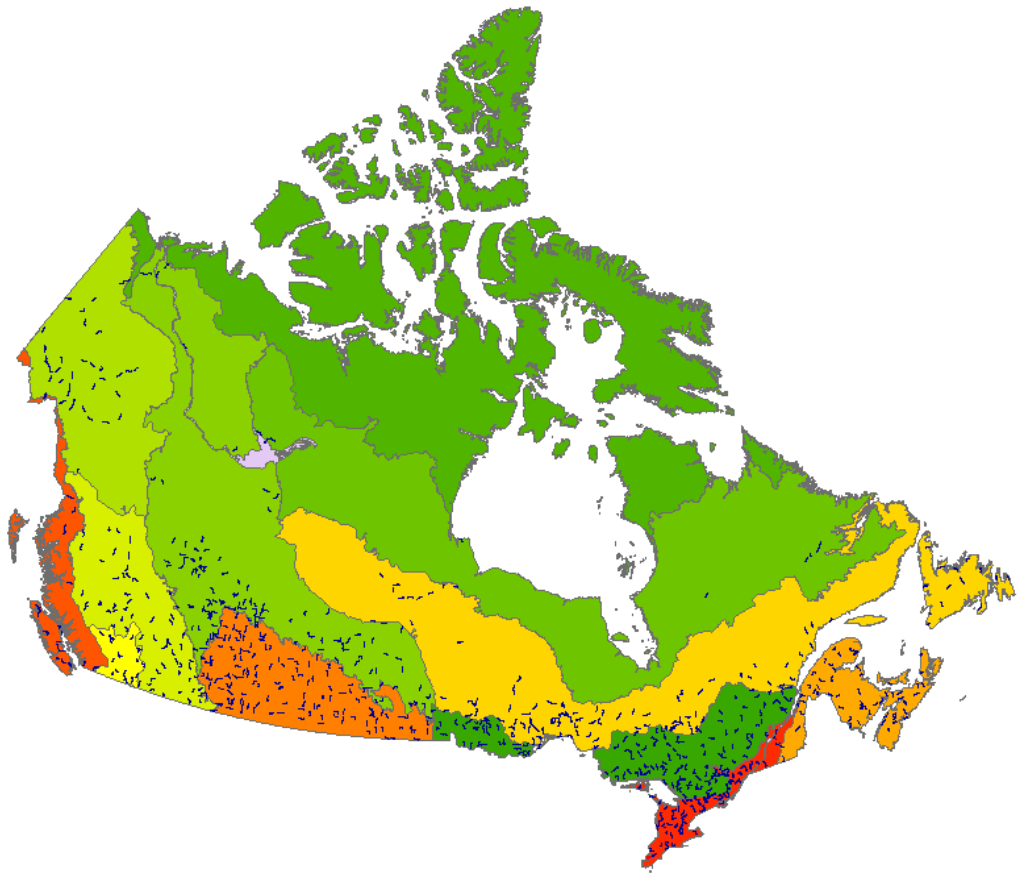


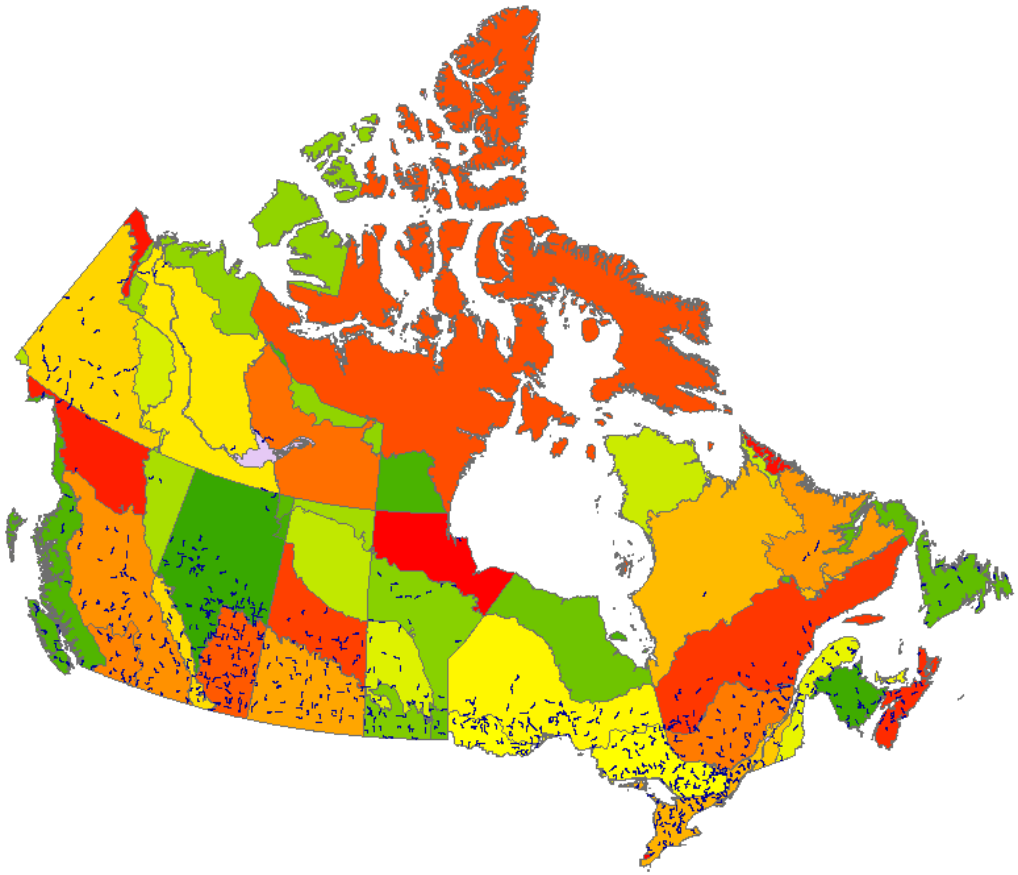

```
mHier_sl<-lme(distance ~ age, random = ~ age | Subject,
Orthodont)
```

One more example: Black-capped Chickadees

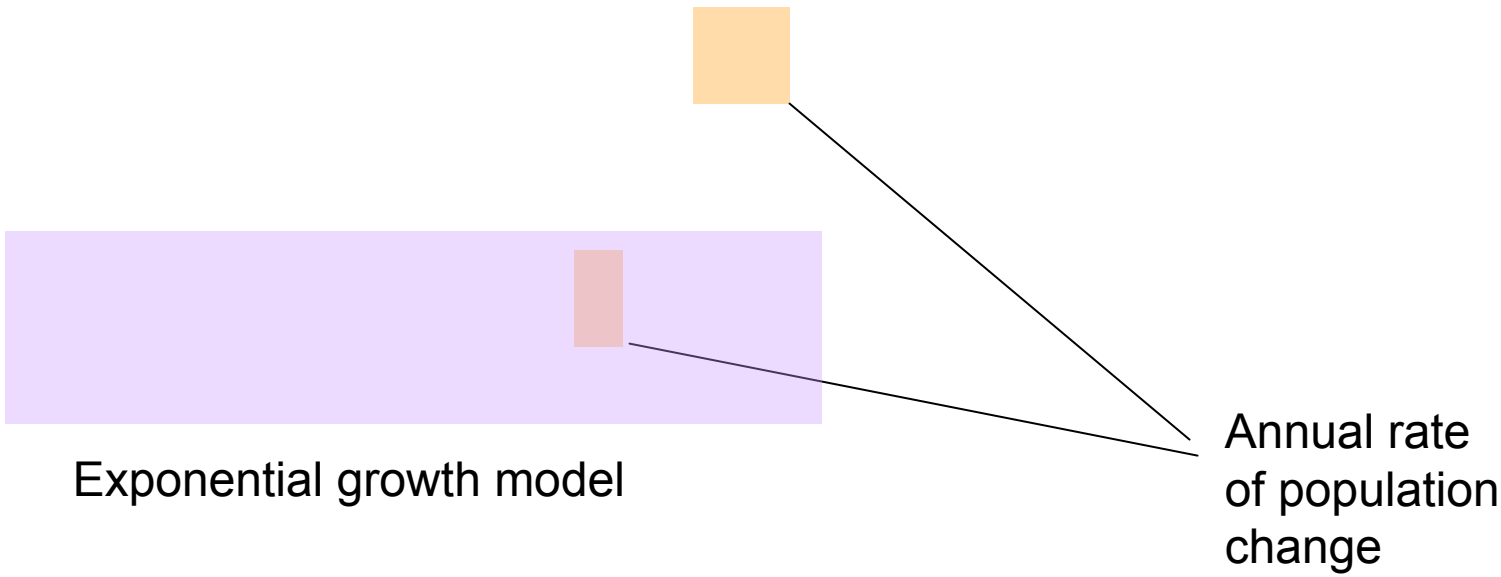
Hierarchical data structure







Poisson population model



Non-hierarchical ("blocked") model

```
m1.fullFix<-glm(Count ~ Date
  + StratumName:Date
  + StratumName
  + Observer
  + firstYear,
  data = testBird,
  family=quasipoisson
  )
```

0.3% per year decline

```
> coef(m1.fullFix)
                                    (Intercept)
                                     0.9092048557
                                             Date
                                    -0.0035591498
StratumNameCA-AB-NORTHERN_ROCKIES
                                     0.4781011718
StratumNameCA-AB-PRAIRIE_POTHOLES
                                    -0.4109260079
StratumNameCA-BC-BOREAL_TAIGA_PLAINS
                                    -8.8370734569
StratumNameCA-BC-GREAT_BASIN
                                    -0.2389701433
```

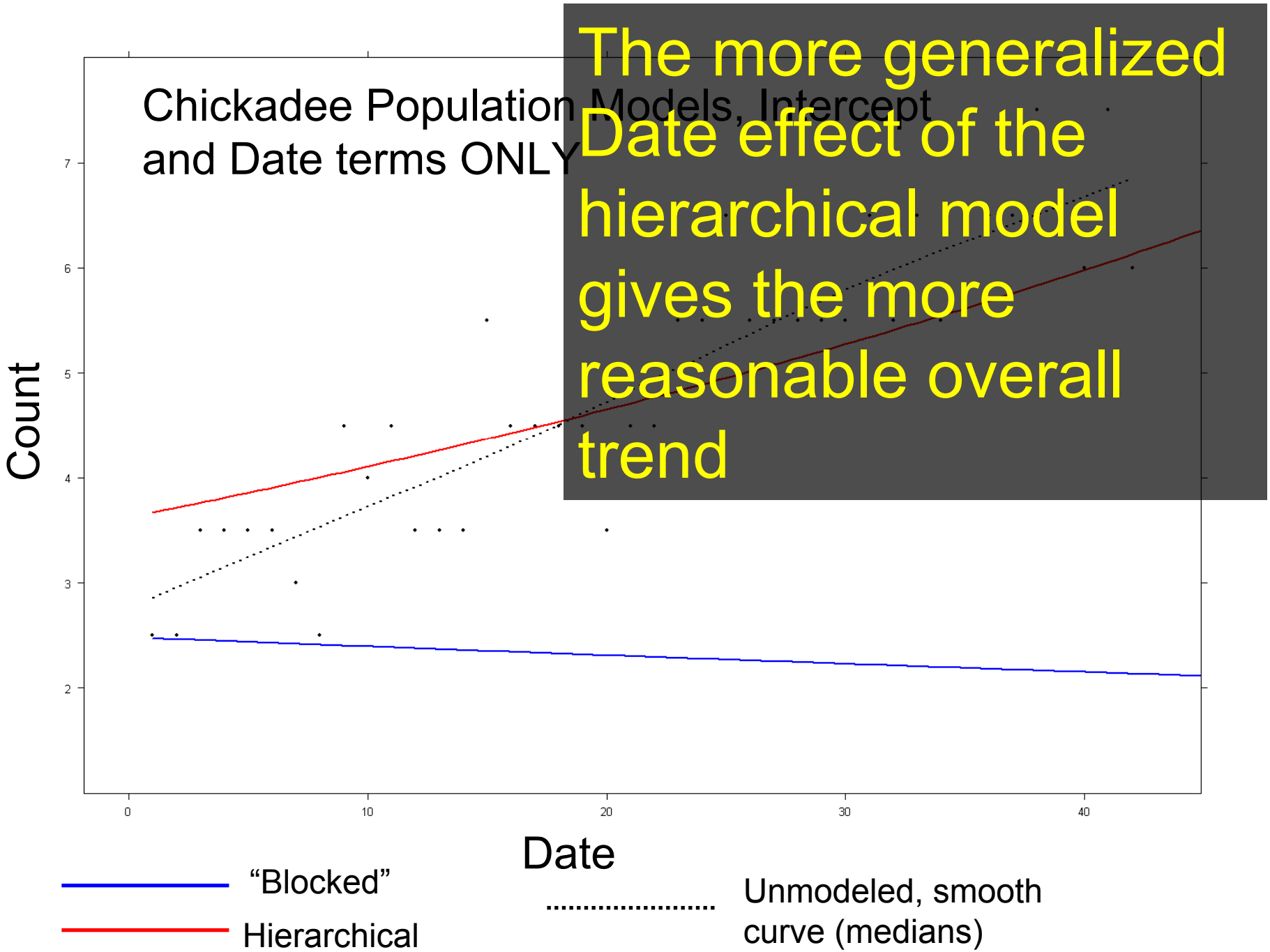
More than one hundred parameters!

Hierarchical model

```
m2.full<-glmmPQL(Count ~ Date,
family = quasipoisson,
  random = list(
    ~Date | StratumName,
    ~ 1 | Observer,
    ~ 1 | firstYear
    ),
  data = testBird
  )
```

1.3% per year increase!

```
> fixef(m2.full)
(Intercept)        Date
 1.28710348  0.01253125
```


Chickadee Population Models, Intercept and Date terms ONLY
The more generalized Date effect of the hierarchical model gives the more reasonable overall trend
Count
Date
0
10
20
30
40
2
3
4
5
6
7
"Blocked"
Hierarchical
Unmodeled, smooth curve (medians)

Further Reading

1. Gelman and Hill (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models. New York: Cambridge University Press.

2. Pinheiro and Bates (2000). Mixed-effects models in S and S-PLUS. New York: Springer.