

Outline

Mixed models in R using the lme4 package

Part 6: Interactions

Douglas Bates

University of Wisconsin - Madison
and R Development Core Team
<Douglas.Bates@R-project.org>

Max Planck Institute for Ornithology
Seewiesen
July 21, 2009

Interactions with grouping factors

The Machines data

Scalar interactions or vector-valued random effects?

The brain activation data

Considering differences

Fixed-effects for the animals

Summary

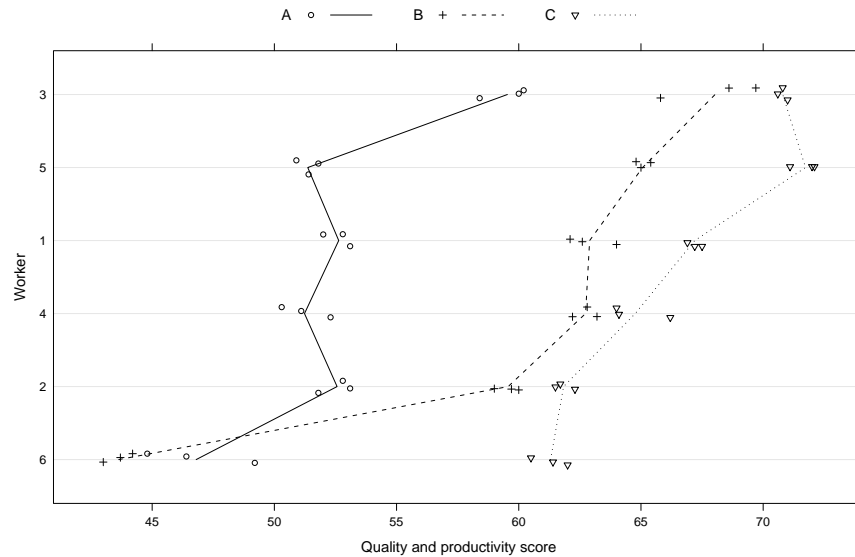
Interactions of covariates and grouping factors

- ▶ For longitudinal data, having a random effect for the slope w.r.t. time by subject is reasonably easy to understand.
- ▶ Although not generally presented in this way, these random effects are an interaction term between the grouping factor for the random effect (`Subject`) and the time covariate.
- ▶ We can also define interactions between a categorical covariate and a random-effects grouping factor.
- ▶ Different ways of expressing such interactions lead to different numbers of random effects. These different definitions have different levels of complexity, affecting both their expressive power and the ability to estimate all the parameters in the model.

Machines data

- ▶ Milliken and Johnson (1989) provide (probably artificial) data on an experiment to measure productivity according to the machine being used for a particular operation.
- ▶ In the experiment, a sample of six different operators used each of the three machines on three occasions — a total of nine runs per operator.
- ▶ These three machines were the specific machines of interest and we model their effect as a fixed-effect term.
- ▶ The operators represented a sample from the population of potential operators. We model this factor, (`Worker`), as a random effect.
- ▶ This is a replicated “subject/stimulus” design with a fixed set of stimuli that are themselves of interest. (In other situations the stimuli may be a sample from a population of stimuli.)

Machines data plot



Comments on the data plot

- ▶ There are obvious differences between the scores on different machines.
- ▶ It seems likely that `Worker` will be a significant random effect, especially when considering the low variation within replicates.
- ▶ There also appears to be a significant `Worker:Machine` interaction. `Worker 6` has a very different pattern w.r.t. machines than do the others.
- ▶ We can approach the interaction in one of two ways: define simple, scalar random effects for `Worker` and for the `Worker:Machine` interaction or define vector-valued random effects for `Worker`

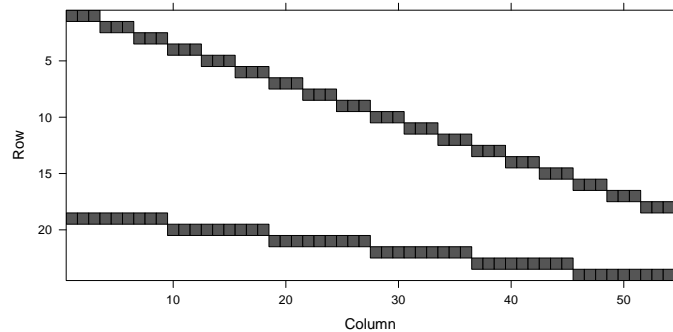
Random effects for subject and subject:stimulus

```
Linear mixed model fit by REML
Formula: score ~ Machine + (1 | Worker) + (1 | Worker:Machine)
Data: Machines
AIC   BIC logLik deviance REMLdev
227.7 239.6 -107.8   225.5   215.7
Random effects:
Groups      Name      Variance Std.Dev.
Worker:Machine (Intercept) 13.90946 3.72954
Worker      (Intercept) 22.85849 4.78105
Residual                    0.92463 0.96158
Number of obs: 54, groups: Worker:Machine, 18; Worker, 6
Fixed effects:
              Estimate Std. Error t value
(Intercept)   52.356     2.486   21.063
MachineB       7.967     2.177    3.660
MachineC      13.917     2.177    6.393
```

Characteristics of the scalar interaction model

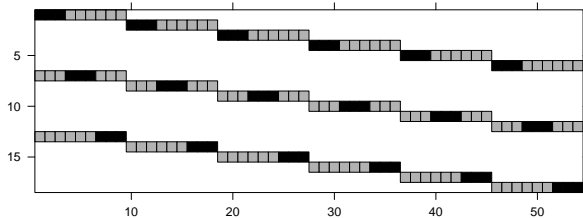
- ▶ The model incorporates simple, scalar random effects for `Worker` and for the `Worker:Machine` interaction.
- ▶ These two scalar random-effects terms have $q_1 = q_2 = 1$ so they contribute $n_1 = 6$ and $n_2 = 18$ random effects for a total of $q = 24$. There are 2 variance-component parameters.
- ▶ The random effects allow for an overall shift in level for each worker and a separate shift for each combination of worker and machine. The unconditional distributions of these random effects are independent. The unconditional variances of the interaction random effects are constant.
- ▶ The main restriction in this model is the assumption of constant variance and independence of the interaction random effects.

Model matrix Z^T for the scalar interaction model



- ▶ Because we know these are scalar random effects we can recognize the pattern of a balanced, nested, two-factor design, similar to that of the model for the Pastes data.

Characteristics of the vector-valued r.e. model



- ▶ We use the specification $(0 + \text{Machine} | \text{Worker})$ to force an “indicator” parameterization of the random effects.
- ▶ In this image the 1's are black. The gray positions are non-systematic zeros (initially zero but can become nonzero).
- ▶ Here $k = 1$, $q_1 = 3$ and $n_1 = 6$ so we have $q = 18$ random effects but $q_1(q_1 + 1)/2 = 6$ variance-component parameters to estimate.

Vector-valued random effects by subject

```

Linear mixed model fit by REML
Formula: score ~ Machine + (0 + Machine | Worker)
Data: Machines
   AIC   BIC logLik deviance REMLdev
228.3 248.2 -104.2   216.6   208.3
Random effects:
Groups   Name      Variance Std.Dev. Corr
Worker  MachineA 16.64097 4.07933
        MachineB 74.39556 8.62529 0.803
        MachineC 19.26756 4.38948 0.623 0.771
Residual              0.92463 0.96158
Number of obs: 54, groups: Worker, 6
Fixed effects:
              Estimate Std. Error t value
(Intercept)   52.356      1.681   31.150
MachineB       7.967      2.421    3.291
MachineC      13.917      1.540    9.037
    
```

Comparing the model fits

- ▶ Although not obvious from the specifications, these model fits are nested. If the variance-covariance matrix for the vector-valued random effects has a special form, called *compound symmetry*, the model reduces to model fm1.
- ▶ The p-value of 6.5% may or may not be significant.

```

> fm2M <- update(fm2, REML = FALSE)
> fm1M <- update(fm1, REML = FALSE)
> anova(fm2M, fm1M)
    
```

```

Data: Machines
Models:
fm1M: score ~ Machine + (1 | Worker) + (1 | Worker:Machine)
fm2M: score ~ Machine + (0 + Machine | Worker)
   Df    AIC    BIC logLik Chisq Chi Df Pr(>Chisq)
fm1M  6  237.27  249.20 -112.64
fm2M 10  236.42  256.31 -108.21  8.8516     4   0.06492
    
```

Model comparisons eliminating the unusual combination

- ▶ In a case like this we may want to check if a single, unusual combination (Worker 6 using Machine "B") causes the more complex model to appear necessary. We eliminate that unusual combination.

```
> Machines1 <- subset(Machines, !(Worker == "6" & Machine ==  
+ "B"))  
> xtabs(~Machine + Worker, Machines1)
```

```
      Worker  
Machine 1 2 3 4 5 6  
  A 3 3 3 3 3 3  
  B 3 3 3 3 3 0  
  C 3 3 3 3 3 3
```

Model comparisons without the unusual combination

```
> fm1aM <- lmer(score ~ Machine + (1 | Worker) + (1 |  
+ Worker:Machine), Machines1, REML = FALSE)  
> fm2aM <- lmer(score ~ Machine + (0 + Machine | Worker),  
+ Machines1, REML = FALSE)  
> anova(fm2aM, fm1aM)
```

Data: Machines1

Models:

fm1aM: score ~ Machine + (1 | Worker) + (1 | Worker:Machine)

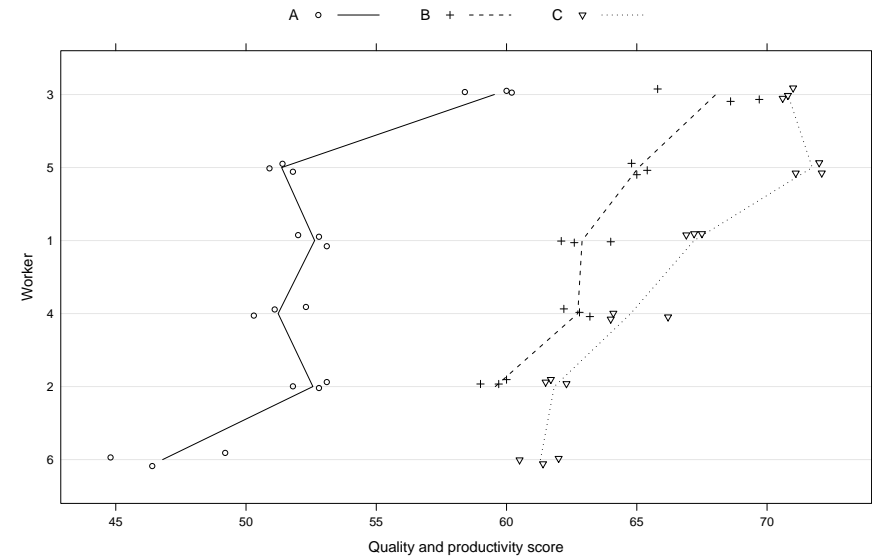
fm2aM: score ~ Machine + (0 + Machine | Worker)

| | Df | AIC | BIC | logLik | Chisq | Chi Df | Pr(>Chisq) |
|--|----|-----|-----|--------|-------|--------|------------|
|--|----|-----|-----|--------|-------|--------|------------|

| | | | | | | | |
|-------|---|---------|---------|---------|--|--|--|
| fm1aM | 6 | 208.554 | 220.145 | -98.277 | | | |
|-------|---|---------|---------|---------|--|--|--|

| | | | | | | | |
|-------|----|---------|---------|---------|--------|---|---------|
| fm2aM | 10 | 208.289 | 227.607 | -94.144 | 8.2655 | 4 | 0.08232 |
|-------|----|---------|---------|---------|--------|---|---------|

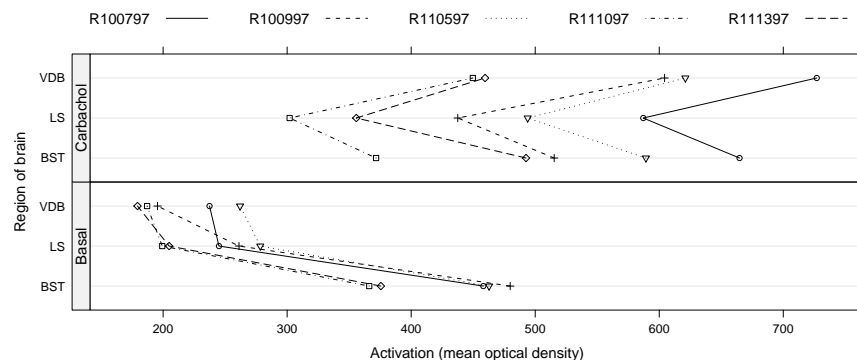
Machines data after eliminating the unusual combination



Trade-offs when defining interactions

- ▶ It is important to realize that estimating scale parameters (i.e. variances and covariances) is considerably more difficult than estimating location parameters (i.e. means or fixed-effects coefficients).
- ▶ A vector-valued random effect term having q_i random effects per level of the grouping factor requires $q_i(q_i + 1)/2$ variance-covariance parameters to be estimated. A simple, scalar random effect for the interaction of a “random-effects” factor and a “fixed-effects” factor requires only 1 additional variance-covariance parameter.
- ▶ Especially when the “fixed-effects” factor has a moderate to large number of levels, the trade-off in model complexity argues against the vector-valued approach.
- ▶ One of the major sources of difficulty in using the `lme4` package is the tendency to overspecify the number of random effects per level of a grouping factor.

Brain activation data from West, Welch and Gałeckı (2007)



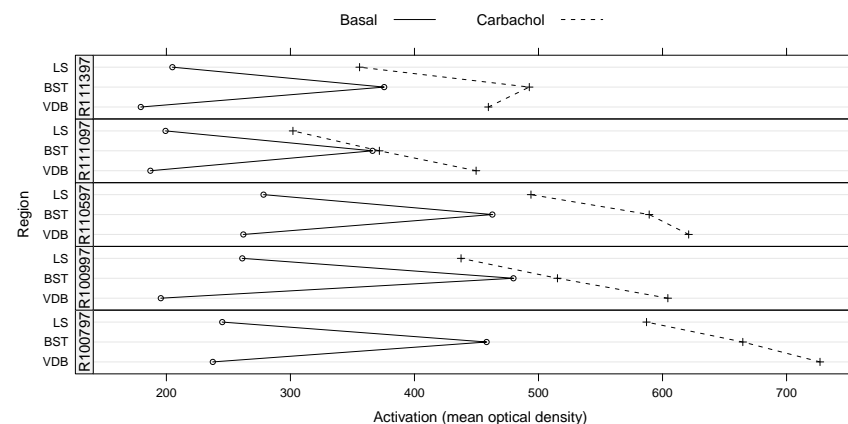
- ▶ In the experiment seven different regions of five rats' brains were imaged in a basal condition (after injection with saline solution) and after treatment with the drug Carbachol. The data provided are from three regions.
- ▶ This representation of the data is similar to the figure on the cover of West, Welch and Gałeckı (2007).

Reproducing the models from West et al.

- ▶ These data are analyzed in West et al. (2007) allowing for main effects for treatment and region, a fixed-effects interaction of these two factors and vector-valued random effects for the intercept and the treatment by animal.
- ▶ Note that this will require estimating three variance component parameters from data on five animals.
- ▶ Their final model also allowed for different residual variances by treatment. We won't discuss that here.
- ▶ We choose the order of the levels of region to produce the same parameterization of the fixed effects.

```
'data.frame': 30 obs. of 4 variables:
 $ animal : Factor w/ 5 levels "R100797","R100997",...: 4 4 4 4 4 5
 $ treatment: Factor w/ 2 levels "Basal","Carbachol": 1 1 1 2 2 2 1 1 1
 $ region : Factor w/ 3 levels "VDB","BST","LS": 2 3 1 2 3 1 2 3 1 2
 $ activate : num 366 199 187 372 302 ...
```

Brain activation data in an alternative layout



- ▶ The animals have similar patterns of changes but different magnitudes.

Model 5.1 from West et al.

```
Linear mixed model fit by REML
Formula: activate ~ region * treatment + (1 | animal)
Data: ratbrain
AIC BIC logLik deviance REMLdev
291.3 302.5 -137.6 325.3 275.3
Random effects:
Groups Name Variance Std.Dev.
animal (Intercept) 4849.8 69.64
Residual 2450.3 49.50
Number of obs: 30, groups: animal, 5
Fixed effects:

```

| | Estimate | Std. Error | t value |
|------------------------------|----------|------------|---------|
| (Intercept) | 212.29 | 38.21 | 5.556 |
| regionBST | 216.21 | 31.31 | 6.906 |
| regionLS | 25.45 | 31.31 | 0.813 |
| treatmentCarbachol | 360.03 | 31.31 | 11.500 |
| regionBST:treatmentCarbachol | -261.82 | 44.27 | -5.914 |
| regionLS:treatmentCarbachol | -162.50 | 44.27 | -3.670 |

Model 5.2 from West et al.

```
Linear mixed model fit by REML
Formula: activate ~ region * treatment + (treatment | animal)
Data: ratbrain
AIC BIC logLik deviance REMLdev
269.2 283.2 -124.6 292.7 249.2
Random effects:
Groups Name Variance Std.Dev. Corr
animal (Intercept) 1284.3 35.837
 treatmentCarbachol 6371.3 79.821 0.801
Residual 538.9 23.214
Number of obs: 30, groups: animal, 5
Fixed effects:
```

| | Estimate | Std. Error | t value |
|------------------------------|----------|------------|---------|
| (Intercept) | 212.29 | 19.10 | 11.117 |
| regionBST | 216.21 | 14.68 | 14.726 |
| regionLS | 25.45 | 14.68 | 1.733 |
| treatmentCarbachol | 360.03 | 38.60 | 9.328 |
| regionBST:treatmentCarbachol | -261.82 | 20.76 | -12.610 |
| regionLS:treatmentCarbachol | -162.50 | 20.76 | -7.826 |

A variation on model 5.2 from West et al.

```
Linear mixed model fit by REML
Formula: activate ~ region * treatment + (0 + treatment | animal)
Data: ratbrain
AIC BIC logLik deviance REMLdev
269.2 283.2 -124.6 292.7 249.2
Random effects:
Groups Name Variance Std.Dev. Corr
animal treatmentBasal 1284.3 35.837
 treatmentCarbachol 12238.1 110.626 0.902
Residual 538.9 23.214
Number of obs: 30, groups: animal, 5
Fixed effects:
```

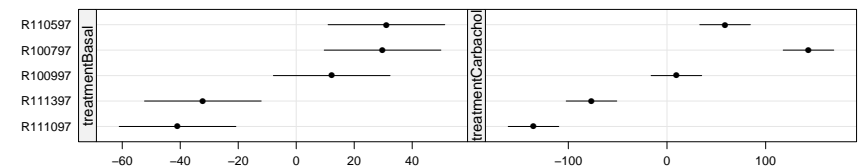
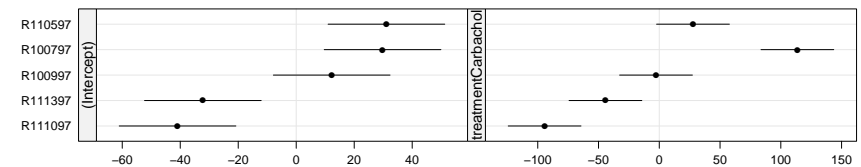
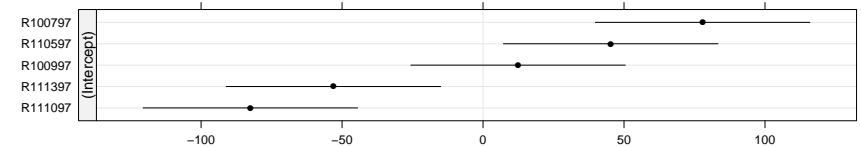
| | Estimate | Std. Error | t value |
|------------------------------|----------|------------|---------|
| (Intercept) | 212.29 | 19.10 | 11.117 |
| regionBST | 216.21 | 14.68 | 14.726 |
| regionLS | 25.45 | 14.68 | 1.733 |
| treatmentCarbachol | 360.03 | 38.60 | 9.328 |
| regionBST:treatmentCarbachol | -261.82 | 20.76 | -12.610 |
| regionLS:treatmentCarbachol | -162.50 | 20.76 | -7.826 |

Simple scalar random effects for the interaction

```
Linear mixed model fit by REML
Formula: activate ~ region * treatment + (1 | animal) + (1 | animal:treatment)
Data: ratbrain
AIC BIC logLik deviance REMLdev
274.7 287.3 -128.4 302.1 256.7
Random effects:
Groups Name Variance Std.Dev.
animal:treatment (Intercept) 3185.7 56.442
animal (Intercept) 3575.5 59.796
Residual 538.9 23.214
Number of obs: 30, groups: animal:treatment, 10; animal, 5
Fixed effects:
```

| | Estimate | Std. Error | t value |
|------------------------------|----------|------------|---------|
| (Intercept) | 212.29 | 38.21 | 5.556 |
| regionBST | 216.21 | 14.68 | 14.726 |
| regionLS | 25.45 | 14.68 | 1.733 |
| treatmentCarbachol | 360.03 | 38.60 | 9.328 |
| regionBST:treatmentCarbachol | -261.82 | 20.76 | -12.610 |
| regionLS:treatmentCarbachol | -162.50 | 20.76 | -7.826 |

Prediction intervals for the random effects



Is this “overmodeling” the data?

- ▶ The prediction intervals for the random effects indicate that the vector-valued random effects are useful, as does a model comparison.

```
Data: ratbrain
```

```
Models:
```

```
m51M: activate ~ region * treatment + (1 | animal)
m52M: activate ~ region * treatment + (treatment | animal)
      Df    AIC    BIC logLik Chisq Chi Df Pr(>Chisq)
m51M  8  341.34  352.55 -162.67
m52M 10  312.72  326.73 -146.36 32.615      2 8.276e-08
```

- ▶ However, these models incorporate many fixed-effects parameters and random effects in a model of a relatively small amount of data. Is this too much?
- ▶ There are several ways we can approach this:
 - ▶ Simplify the model by considering the difference in activation under the two conditions within the same animal:region combination (i.e. approach it like a paired t-test).
 - ▶ Model the five animals with fixed effects and use F-tests.
 - ▶ Assess the precision of the variance estimates (done later).

Checking for balanced and unreplicated; tabling activate

- ▶ We saw the balance in the data plots but we can check too
 - > `fable(xtabs(~treatment + region + animal, ratbrain))`

```
          animal R100797 R100997 R110597 R111097 R111397
treatment region
Basal     VDB           1         1         1         1         1
          BST           1         1         1         1         1
          LS           1         1         1         1         1
Carbachol VDB           1         1         1         1         1
          BST           1         1         1         1         1
          LS           1         1         1         1         1
```

- ▶ In `xtabs` we can use a two-sided formula to tabulate a variable
 - > `fable(atab <- xtabs(activate ~ treatment + animal + region, ratbrain))`

```
          region    VDB    BST    LS
treatment animal
Basal     R100797    237.42 458.16 245.04
          R100997    195.51 479.81 261.19
          R110597    262.05 462.79 278.33
          R111097    187.11 366.19 199.31
          R111397    179.38 375.58 204.85
Carbachol R100797    726.96 664.72 587.10
          R100997    604.29 515.29 437.56
```

Considering differences

- ▶ Before we can analyze the differences at each animal:region combination we must first calculate them.
- ▶ We could do this by subsetting the `ratbrain` data frame for the "Basal" and "Carbachol" levels of the treatment factor and forming the difference of the two activate columns. For this to be correct we must have the same ordering of levels of the animal and region factors in each half. It turns out we do but we shouldn't count on this (remember “Murphy's Law”?).
- ▶ A better approach is to reshape the data frame (but this is complicated) or to use `xtabs` to align the levels. First we should check that the data are indeed balanced and unreplicated.

Taking differences

- ▶ The `atab` object is an array with additional attributes

```
xtabs [1:2, 1:5, 1:3] 237 727 196 604 262 ...
- attr(*, "dimnames")=List of 3
  ..$ treatment: chr [1:2] "Basal" "Carbachol"
  ..$ animal   : chr [1:5] "R100797" "R100997" "R110597" "R111097"
  ..$ region   : chr [1:3] "VDB" "BST" "LS"
- attr(*, "class")= chr [1:2] "xtabs" "table"
- attr(*, "call")= language xtabs(formula = activate ~ treatment + animal + region, data = ratbrain)
```
- ▶ Use `apply` to take differences over dimension 1
 - > `(diffs <- as.table(apply(atab, 2:3, diff)))`

```
          region
animal    VDB    BST    LS
R100797  489.54 206.56 342.06
R100997  408.78  35.48 176.37
R110597  359.02 126.46 215.60
R111097  262.59   5.52 102.71
R111397  280.20 117.00 150.89
```

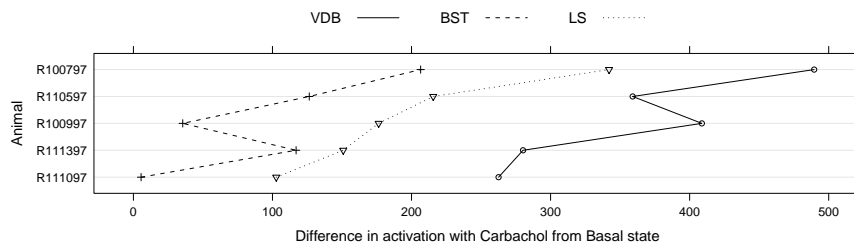
Taking differences (cont'd)

- ▶ Finally, convert the table of differences to a data frame.

```
> str(diffs <- as.data.frame(diffs))
```

```
'data.frame': 15 obs. of 3 variables:
 $ animal: Factor w/ 5 levels "R100797","R100997",...: 1 2 3 4 5 1 2 3 4
 $ region: Factor w/ 3 levels "VDB","BST","LS": 1 1 1 1 1 2 2 2 2 2 ...
 $ Freq : num 490 409 359 263 280 ...
```

```
> names(diffs)[3] <- "actdiff"
```



Using fixed-effects for the animals

- ▶ There are five experimental units (animals) in this study. That is about the lower limit under which we could hope to estimate variance components.
- ▶ We should compare with a fixed-effects model.
- ▶ If we wish to evaluate coefficients for treatment or region we must be careful about the “contrasts” that are used to create the model. However, the analysis of variance table does not depend on the contrasts.
- ▶ We use `aov` to fit the fixed-effects model so that a summary is the analysis of variance table.
- ▶ The fixed-effects anova table is the sequential table with main effects first, then two-factor interactions, etc. The anova table for an `lmer` model gives the contributions of the fixed-effects after removing the contribution of the random effects, which include the `animal:treatment` interaction in model `m52`.

A model for the differences

```
Linear mixed model fit by REML
Formula: actdiff ~ region + (1 | animal)
Data: diffs
AIC BIC logLik deviance REMLdev
147.4 150.9 -68.68 162.3 137.4
Random effects:
Groups Name Variance Std.Dev.
animal (Intercept) 6209.9 78.803
Residual 1562.2 39.524
Number of obs: 15, groups: animal, 5
Fixed effects:
```

| | Estimate | Std. Error | t value |
|-------------|----------|------------|---------|
| (Intercept) | 360.03 | 39.42 | 9.132 |
| regionBST | -261.82 | 25.00 | -10.474 |
| regionLS | -162.50 | 25.00 | -6.501 |

```
Correlation of Fixed Effects:
(Intr) rgnBST
regionBST -0.317
regionLS -0.317 0.500
```

Fixed-effects anova versus random effects

```
> summary(m52f <- aov(activate ~ animal * treatment +
+ region * treatment, ratbrain))
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|------------------|----|--------|---------|---------|-----------|
| animal | 4 | 126197 | 31549 | 58.544 | 2.376e-09 |
| treatment | 1 | 358347 | 358347 | 664.965 | 1.844e-14 |
| region | 2 | 100998 | 50499 | 93.708 | 1.465e-09 |
| animal:treatment | 4 | 40384 | 10096 | 18.734 | 6.973e-06 |
| treatment:region | 2 | 87352 | 43676 | 81.047 | 4.244e-09 |
| Residuals | 16 | 8622 | 539 | | |

```
> anova(m52)
```

```
Analysis of Variance Table
Df Sum Sq Mean Sq F value
region 2 100998 50499 93.708
treatment 1 18900 18900 35.072
region:treatment 2 87352 43676 81.047
```

- ▶ Except for the treatment factor, the anova tables are nearly identical.

Summary

- ▶ It is possible to fit complex models to balanced data sets from carefully designed experiments but one should always be cautious of creating a model that is too complex.
- ▶ I prefer to proceed incrementally, taking time to examine data plots, rather than starting with a model incorporating all possible terms.
- ▶ Some feel that one should be able to specify the analysis (and, in particular, the analysis of variance table) before even beginning to collect data. I am more of a model-builder and try to avoid dogmatic approaches.
- ▶ For the `ratbrain` data I would be very tempted to take differences and analyze it as a randomized blocked design.