PSYCH 6140A: Multivariate Analysis
Module 10: Analyzing within-subjects experimental data using multilevel modeling

We can view data from a within-subjects design as hierarchically nested such that repeated observations represent Level 1 of the data structure, while individual subjects or participants represent Level 2.
That is, repeated measures are nested, or clustered, within subjects.

Therefore, multilevel models (MLMs) can be fitted to data from studies which have a within-subjects or repeated-measures design.

The data used in this example were originally reported by Douglas et al. (2004). The aim of their experiment was to examine nucleotide activation (guanine nucleotide bonding) in three different brain nuclei (i.e., brain regions) among five adult male rats across two treatments. Both treatments were administered to each rat.

Douglas, C.L., Demarco, G.J., Baghdoyan, H.A., & Lydic, R. (2004). Pontine and basal forebrain cholinergic interaction: Implications for sleep and breathing. *Respiratory Physiology and Neurobiology, 143,* 251.

Therefore, the study has a 2×3 factorial design, where both ways of the design are within-subjects factors. These data could be analyzed using a traditional repeated-measures ANOVA. Alternatively, various two-level MLMs can be fitted to the data (e.g., random intercepts, random slopes…).

Repeated measures data can be organized in either a "wide" format or a "long" format.

In the wide format, there is only one row of data for each subject or participant, and the repeated observations are represented as separate variables, or columns, in the data matrix.

In the long format, there are multiple rows of data for each subject, with the number of rows per subject equaling the number of repeated observations of that subject. There will then be separate variables, or columns, for each repeated measures factor (i.e., each within-subjects factor is a separate column); for example, there may be a column for "time" or "measurement occasion."

To estimate a MLM for repeated measures data, the data need to be organized in the long format.

In the current example, there are the following four variables:

"animal": Unique identifier for each rat
"treat": Level of drug treatment (0 = Basal, 1 = Carbachol)
-- remember that this is a within-subjects factor.
"region": Brain nucleus (1 = BST, 2 = LS, 0 = VDB)
-- the other within-subjects factor
"activate": Nucleotide activation (the dependent variable)

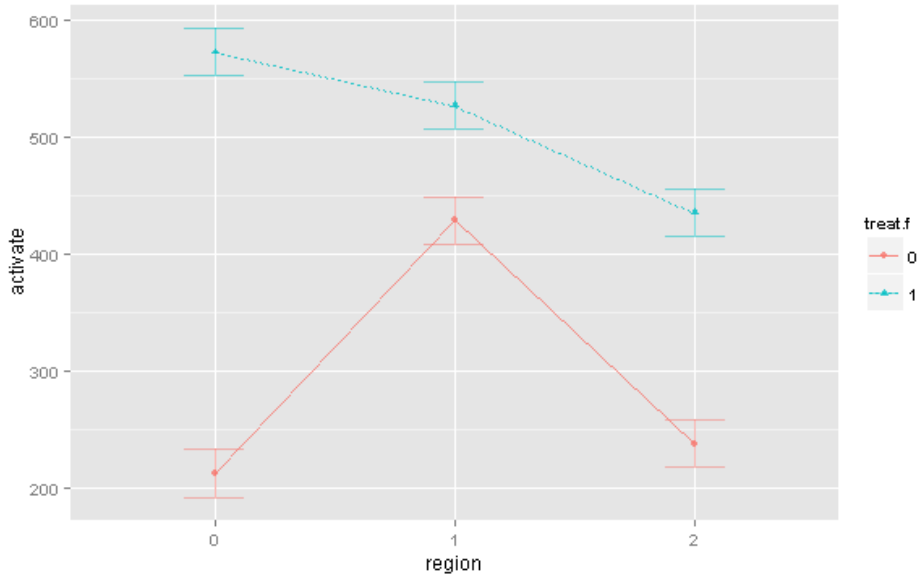
Here is what the data look like for the first two rats (*animal* = 097 and *animal* = 397):

animal	activate	region	treat
097	366.19	1	0
097	199.31	2	0
097	187.11	0	0
097	371.71	1	1
097	302.02	2	1
097	449.7	0	1
397	375.58	1	0
397	204.85	2	0
397	179.38	0	0
397	492.58	1	1
397	355.74	2	1
397	459.58	0	1

There are six repeated measures of the dependent variable *activate* for each rat: *activate* is measured across each of three *region*s across the two levels of *treat*.

Because there are five rats, the complete data set has 5 × 6 = 30 rows of data.

To get a sense of the data, here is a plot of the means of *activate* across the 2 × 3 = 6 cells of the within-subjects design:



To begin, though, we will only model *activate* as a function of *region*, ignoring (for now) the *treat* variable.

Recall that a major idea of the MLM is that there is a separate regression equation for each Level 2 unit, which in this example are the five rats.
Thus, here there are five different regressions of *activate* on *region*, one for each rat.
Keep in mind that each region is observed twice within each rat; thus each of the five regressions is based on only six observations.

Additionally, because *region* is a nominally scaled variable with three levels, we need to represent it with two dichotomous variables.
Here, we will use two dummy variables, *D*1 to compare the VDB region (*region* = 0) to the BST region (*region* = 1) and *D*2 to compare the VDB region (*region* = 0) to the LS region (*region* = 2).

Thus, our Level 1 model for the regression of *activate* on *region* is

$$Y_{ij} = \beta_{0j} + \beta_{1j} D1_{ij} + \beta_{2j} D2_{ij} + e_{ij}$$

with *i* indexing Level 1 units (repeated measures within rats) and *j* indexing Level 2 units (different rats).
In this Level 1 equation, Y_{ij} is the value of *activate* in measurement *i* for rat *j*,
β_{0j} is the intercept for rat *j*,
β_{1j} is the regression slope comparing the VDB region to the BST region within rat *j*,
β_{2j} is the regression slope comparing the VDB region to the LS region within rat *j*,
and e_{ij} is the Level 1 error term, which here is specific to measurement *i* in rat *j*.

Level 1 equation: $Y_{ij} = \beta_{0j} + \beta_{1j} D1_{ij} + \beta_{2j} D2_{ij} + e_{ij}$

Random intercept model for *region*

At first, let's allow only the Level 1 intercept to vary across rats, so the Level 2 equations are

$\beta_{0j} = \gamma_{00} + u_{0j}$
$\beta_{1j} = \gamma_{10}$
and
$\beta_{2j} = \gamma_{20}$

γ_{00} is the mean intercept averaged across rats. Here, because the VDB region is the reference level for the two dummy variables, γ_{00} represents the grand mean activation in the VDB region, averaged over the two VDB measurements of all five rats.

u_{0j} is then the difference between this grand mean and the mean VDB activation of rat *j*.
The variance of u_{0j}, τ_{00}, will represent the variability of the mean VDB activation across the different rats.

Next, γ_{10} is the mean difference between activation in the VDB region and activation in the BST region. This effect is considered constant across all rats because there is no random term for β_{1j}.

Finally, γ_{20} is the mean difference between activation in the VDB region and activation in the LS region. This effect is also considered constant across all rats.

This random-intercepts model is equivalent to the traditional, univariate within-subjects ANOVA model assuming *compound symmetry*.

To estimate the model in R, there are two different approaches to dealing with the fact that *region* is a nominal variable with three levels.

The first is to create dummy variables, as we have learned before:

```
D1 <- (region==1)*1
D2 <- (region==2)*1
```

And then use those dummy variables as the Level 1 predictors in the model:

```
randint.region <- lme(activate ~ D1 + D2, random =~ 1|animal)
```

where the option (`random = ~1|animal`) specifies that intercepts are to vary randomly across *animal* (i.e., rats).

The summary method gives us

```
> summary(randint.region)
Linear mixed-effects model fit by REML
 Data: NULL
       AIC      BIC    logLik
  364.4207 370.8999 -177.2104
Random effects:
 Formula: ~1 | animal
        (Intercept) Residual
StdDev:    40.90697  146.659
Fixed effects: activate ~ D1 + D2
             Value Std.Error DF   t-value p-value
(Intercept) 392.307  49.85542 23  7.868893  0.0000
D1           85.301  65.58791 23  1.300560  0.2063
D2          -55.800  65.58791 23 -0.850767  0.4037
```

Alternatively, we can deal with the *region* variable by creating a new variable that R will explicitly recognize and treat as a discrete, categorical variable (i.e., a *factor*):

```
region.f <- factor(region)
```

and then we use the factor as the sole regressor in the model:

```
randint.region.f <- lme(activate ~ region.f, random = ~1|animal)
```

Now the summary method produces

```
> summary(randint.region.f)
Linear mixed-effects model fit by REML
 Data: NULL
       AIC      BIC    logLik
  364.4207 370.8999 -177.2104

Random effects:
 Formula: ~1 | animal
        (Intercept) Residual
StdDev:    40.90697  146.659

Fixed effects: activate ~ region.f
             Value Std.Error DF   t-value p-value
(Intercept) 392.307  49.85542 23  7.868893  0.0000
region.f1    85.301  65.58791 23  1.300560  0.2063
region.f2   -55.800  65.58791 23 -0.850767  0.4037
```

which is identical to the results we got by explicitly creating the dummy variables ourselves.

(But you have to be careful about knowing which level of the factor variable is being used as the reference category; by default, the reference category will be the level of the factor variable with the lowest value, which here is *region* = 0).

So we see that we can use the single `region.f` factor as a substitute for the two dummy variables in the rest of these analyses.

```
> summary(randint.region.f)
Linear mixed-effects model fit by REML
 Data: NULL
       AIC      BIC    logLik
  364.4207 370.8999 -177.2104

Random effects:
 Formula: ~1 | animal
        (Intercept) Residual
StdDev:    40.90697  146.659

Fixed effects: activate ~ region.f
               Value Std.Error DF   t-value p-value
(Intercept) 392.307  49.85542 23  7.868893  0.0000
region.f1    85.301  65.58791 23  1.300560  0.2063
region.f2   -55.800  65.58791 23 -0.850767  0.4037
```

Level 1 equation: $Y_{ij} = \beta_{0j} + \beta_{1j} D1_{ij} + \beta_{2j} D2_{ij} + e_{ij}$

Level 2 equations:

$\beta_{0j} = \gamma_{00} + u_{0j}$

$\beta_{1j} = \gamma_{10}$

and

$\beta_{2j} = \gamma_{20}$

These results give us $\text{var}(u_{0j}) = \tau_{00} = (40.91)^2$,

$\hat{\gamma}_{00} = 392.31$,

$\hat{\gamma}_{10} = 85.30$, which is not significant, t (23) = 1.30, p = .21,

and $\hat{\gamma}_{20} = -55.80$, which also is not significant, t (23) = 0.85, p = .40.

So overall it appears as if the average levels of activation do not differ very much across the three regions.

We can view this MLM analysis as an alternative to a traditional one-way, within-subjects ANOVA.

But at this point we know that we are ignoring the treatment effect, so we won't take this result too seriously.

Random slopes model for *region*

Level 1 equation: $Y_{ij} = \beta_{0j} + \beta_{1j} D1_{ij} + \beta_{2j} D2_{ij} + e_{ij}$

Next, let's see if the slope terms in the relationship between *region* and *activate* are heterogeneous across the different rats.

So our Level 2 equations continue to include a random term for the intercepts:
$\beta_{0j} = \gamma_{00} + u_{0j}$

But now we also have random terms for the slopes of the two dummy variables:
$\beta_{1j} = \gamma_{10} + u_{1j}$
and
$\beta_{2j} = \gamma_{20} + u_{2j}$

We can fit this model with

```
randslope.region <- lme(activate ~ region.f, random =~ region.f | animal)
```

and the summary is

```
> summary(randslope.region)
Linear mixed-effects model fit by REML
 Data: NULL
       AIC      BIC    logLik
  374.4207 387.3791 -177.2104

Random effects:
 Formula: ~region.f | animal
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev       Corr
(Intercept) 4.090696e+01 (Intr) rgn.f1
region.f1   1.104266e-02 0
region.f2   7.615206e-03 0      0
Residual    1.466590e+02

Fixed effects: activate ~ region.f
              Value Std.Error DF   t-value p-value
(Intercept) 392.307  49.85542 23  7.868893  0.0000
region.f1    85.301  65.58791 23  1.300560  0.2063
region.f2   -55.800  65.58791 23 -0.850766  0.4037
```

Notice above that the standard deviations of the random slopes seem quite small:

$stdev(u_{1j}) = \sqrt{\tau_{11}} = 0.011 =$ 1.104266e-02

$stdev(u_{2j}) = \sqrt{\tau_{22}} = 0.0076 =$ 7.615206e-03

So it is no surprise that the previous random-intercepts model fits the data just as well as the current random-slopes model:

```
> anova(randint.region.f,randslope.region)
                 Model df      AIC      BIC    logLik   Test      L.Ratio
randint.region.f     1  5 364.4207 370.8999 -177.2104
randslope.region     2 10 374.4207 387.3791 -177.2104 1 vs 2 4.302274e-08
                 p-value
randint.region.f
randslope.region       1
```

Incorporating the treatment effect

Recall that, for each of the five rats, the dependent variable *activate* was observed for each *region* within both of the treatment conditions.

Therefore, just as *region* is a Level 1 predictor, *treat* is also a Level 1 predictor.
In other words, both *region* and *treat* are within-subjects factors.

(If one treatment was given only to some rats and the other treatment was given to the other rats, then *treat* would be a between-subjects predictor, that is, a Level 2 predictor, given that rats are the Level 2 units.)

Additionally, it is pretty clear from the plot of means on p. 2 that there is likely an interaction between *region* and *treat*.

Therefore, the Level 1 equation incorporates *region* (using the two dummy variables), *treat*, and their interaction:

$$Y_{ij} = \beta_{0j} + \beta_{1j} D1_{ij} + \beta_{2j} D2_{ij} + \beta_{3j} X_{ij} + \beta_{4j} (X_{ij} * D1_{ij}) + \beta_{5j} (X_{ij} * D2_{ij}) + e_{ij}$$

where X_{ij} is the value of the dichotomous *treat* variable observed during repeated measure *i* for rat *j*.

Note that "repeated measure *i*" refers to one of the six repeated measurements taken for each rat (i.e., based on the 3 × 2 within-subjects design).

Next, at Level 2 we should at least have random intercepts to account for the non-independence of repeated measures within rats:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

But it will not be possible to have a random term for each of the five slopes from the Level 1 equation because there simply is not enough data; for each rat, there is only one observation within each of the six *region* × *treat* cells of the study design. (Technically speaking, the model would be *under-identified*.)

But we could have random terms for the first-order effect of *region* that is marginal to the interaction (i.e., one for each of the dummy variables representing *region*):

$\beta_{1j} = \gamma_{10} + u_{1j}$
and
$\beta_{2j} = \gamma_{20} + u_{2j}$

Or we could instead have a random term for the first-order effect of *treat* that is marginal to the interaction:

$\beta_{3j} = \gamma_{30} + u_{3j}$

From the previous analysis, it does not seem as if having random slopes for *region* is particularly important, so we will instead estimate a model with random slopes for *treat*.

Therefore, the complete set of Level 2 equations will be
$\beta_{0j} = \gamma_{00} + u_{0j}$
$\beta_{1j} = \gamma_{10}$
$\beta_{2j} = \gamma_{20}$
$\beta_{3j} = \gamma_{30} + u_{3j}$
$\beta_{5j} = \gamma_{50}$
$\beta_{6j} = \gamma_{60}$

So the random-effects variance parameters to be estimated are
$\text{var}(u_{0j}) = \tau_{00}$,
$\text{var}(u_{3j}) = \tau_{33}$,
the covariance between u_{0j} and u_{3j}, $\text{cov}(u_{0j}, u_{3j}) = \tau_{03}$,
and, as always, the Level 1 residual variance, $\text{var}(e_{ij}) = \sigma^2$.

We can fit this model with

```
randslope.2way <- lme(activate ~ D1+D2+treat+D1*treat+D2*treat,
                      random =~ treat | animal)
```

or, equivalently,

```
randslope.2way <- lme(activate ~ region.f+treat+region.f*treat,
                      random =~ treat | animal)
```

or, even more succinctly,

```
randslope.2way <- lme(activate ~ region.f*treat, random =~ treat | animal)
```

(the lme function will automatically include all of the necessary first-order marginal effects).

Then the summary method gives us

```
> summary(randslope.2way)
Linear mixed-effects model fit by REML
 Data: NULL
       AIC     BIC    logLik
  269.1904 280.971 -124.5952

Random effects:
 Formula: ~treat | animal
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev   Corr
(Intercept) 35.83696 (Intr)
treat       79.82012 0.801
Residual    23.21432

Fixed effects: activate ~ region.f + treat + region.f * treat
                   Value Std.Error DF    t-value p-value
(Intercept)      212.294  19.09551 20  11.117483  0.0000
region.f1        216.212  14.68203 20  14.726304  0.0000
region.f2         25.450  14.68203 20   1.733412  0.0984
treat            360.026  38.59808 20   9.327561  0.0000
region.f1:treat -261.822  20.76352 20 -12.609710  0.0000
region.f2:treat -162.500  20.76352 20  -7.826225  0.0000
```

with variance components

```
> VarCorr(randslope.2way)
animal = pdLogChol(treat)
            Variance  StdDev   Corr
(Intercept) 1284.2876 35.83696 (Intr)
treat       6371.2509 79.82012 0.801
Residual     538.9048 23.21432
```

First and foremost, we see that the fixed effect estimates for the two interaction terms are both significant. Therefore, any subsequent interpretation of the results must be based on the *region* × *treat* interaction.
In particular, we can adapt the simple slopes method for probing interactions that we applied earlier in the context of regular, fixed-effects multiple regression.

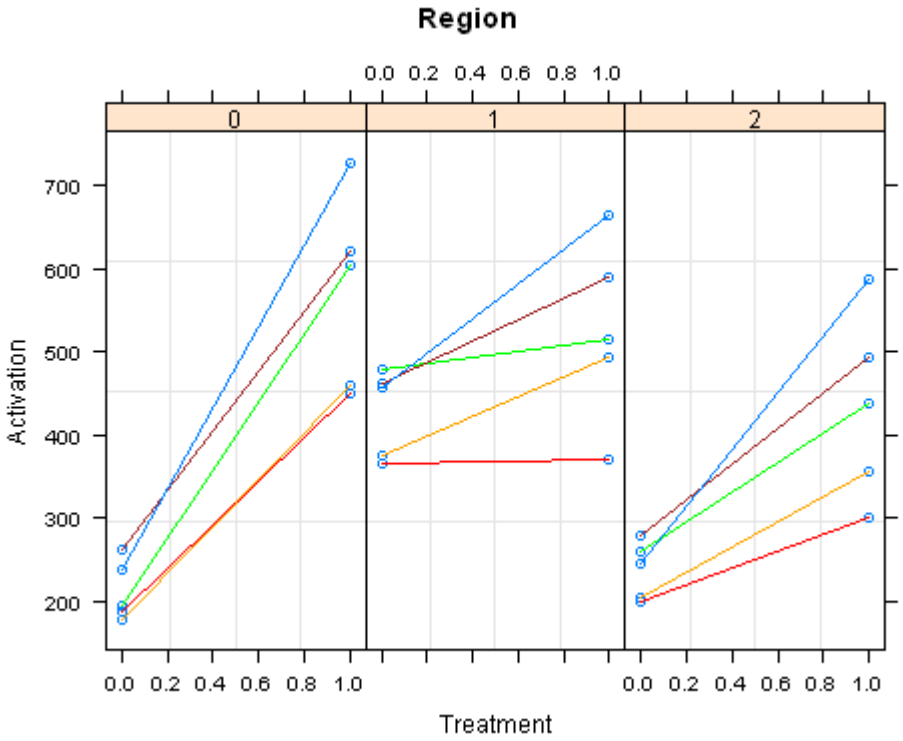
In the output above, the first-order fixed-effect estimate for *treat* is $\hat{\gamma}_{30} = 360.03$. This estimate represents the *mean* difference between the activation levels across the two treatment conditions *within the VDB region*.

This effect is specific to the VDB region because that region is the reference category for the dummy coding system used to represent the three levels of *region*.

Additionally, because the model has a random term for *treat*, this effect is a mean effect across the five rats. The estimated variance of this *treat* effect is $\hat{\tau}_{33} = 6371.25$; this estimate quantifies the heterogeneity of the treatment differences across rats (i.e., the amount that the slope β_{3j} varies across the three rats).

To complete our interpretation of these results, we can continue to probe the interaction by changing the reference category of *region* and re-estimating the model.

But because there are only five rats, we can visualize all of the data simultaneously to get a sense of how the activation dependent variable varies across regions and treatment.
In the plot below, within each region there is a separate line showing the treatment difference for each rat:



From the figure, it seems that the treatment effect is strongest in *region* = 0 (VDB), and perhaps weakest in *region* = 1 (BST).

(The plot displays the actual observed data, and not the predicted values from the multilevel model.)

(correlation structure?)