Repeated Measures ANOVA Multivariate ANOVA and Their Relationship to Linear Mixed Models

EPSY 905: Fundamentals of Multivariate Modeling

Online Lecture #16

Today's Class

- Repeated measures versions of linear models
  - How RM ANOVA models are mixed models with varying assumptions
- Multivariate ANOVA
  - How MANOVA models are mixed models with varying assumptions
- An introduction to multilevel models
- For all today: Models assume data is complete
  - Every observation is recorded :: no missing data

Example Data

- A health researcher is interested in examining the impact of dietary habits and exercise on pulse rate
- A sample of 18 participants is collected
  - Diet factor (BETWEEN SUBJECTS):
    - Nine are vegetarians
    - Nine are omnivores
  - Exercise factor (BETWEEN SUBJECTS) with random assignment:
    - Aerobic stair climbing
    - Racquetball
    - Weight training
  - Three pulse rates (WITHIN SUBJECTS):
    - After warm-up
    - After jogging
    - After running

REPEATED MEASURES ANOVA

Repeated Measures

- Instead of focusing on a single **dependent variable**, we can focus on all three
  - Repeated measures analysis

- In repeated measures GLM, the effects of interest are:
  - Between subjects effects
    - Differences between Diet and Exercise Type
  - Within subjects effects
    - Differences between when pulse rate was taken
    - Not usually a consideration in Multivariate GLM (MANOVA)
  - Interactions between within and between subjects effects

- As we will see, repeated measures GLM uses multivariate data and multivariate distributions

Repeated Measures Distributional Setup

- Because we have repeated observations, we now have a set of variables to predict instead of just one
  - Our example data set had three response variables per person
  - More generally, we will have p response variables per person

$$\mathbf{y}_i = [y_{i1} \quad y_{i2} \quad \cdots \quad y_{ip}]$$

- The GLM is extended to model **multivariate** outcomes

$$f(\mathbf{y}_i|\mathbf{X}_i) \sim N_p(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V})$$

- Now, $\boldsymbol{\beta}$ is size $(1 + k)\ x\ p$

- LS Estimates: $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$
  - Where $\mathbf{Y}$ is size $N\ x\ p$

Unpacking the Repeated Measures Design

- The distributional assumptions are very similar

$$f(\mathbf{y}_i|\mathbf{X}_i) \sim N_p(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V})$$

- **MODEL FOR THE MEANS:**
  - The model for the means is the same (only has more terms)
  - IVs $\mathbf{X}_i$ and linear model weights $\boldsymbol{\beta}$ provide predicted values for each observation

- **MODEL FOR THE VARIANCES:**
  - Because we have repeated observations, the model for the variances is now a covariance matrix of size *p* x *p*
    - <u>Diagonal elements:</u> variance of error for each outcome
    - <u>Off-diagonal elements:</u> covariance of errors for pairs of outcomes

More on Variances

- The key to repeated measures and multivariate ANOVA is the covariance matrix $\mathbf{V}$ structure
  - Type of model dictates types of options for covariance matrix
  - The classical statistics discussed today are one of the three
    - More modern approaches make this more flexible

Three Structures of Classical GLM

- Three structures:

1. Independence: $\mathbf{V} = \sigma_e^2 \mathbf{I}_p$
   - Models if all observations as if they came from separate people
   - No more statistical parameters than original GLM approach
   - Don't use: shown for baseline purposes

2. Repeated Measures: Assumes sphericity of observations
   - Sphericity is a condition that is more strictly enforced by compound symmetry of $\mathbf{V}$ having two parameters:
   - Sphericity is compound symmetry of pairwise differences
     - Diagonal elements: same variance
     - Off-diagonal elements: same covariance
   - No sphericity? Adjustments to F tests

3. Multivariate ANOVA/GLM: Assumes nothing – estimates everything
   - Every unique element in $\mathbf{V}$ is modeled
   - Need more power (i.e., sample size) to make work well
   - Most general procedure

Working Example

- We return to our example, only this time to build the repeated measures and multivariate versions

- We will begin with the unconditional model ("empty model")
  - Provided as a baseline to show how repeated measures works, conceptually – you likely would never run this model

- To run this model, I converted our data from wide format (all data for one observation per row) to long format (only one observation per column)

Example Syntax: Data Transformation

mv16epsy905_lecture12.R × | long_data × | data01 ×

Filter

	exertype	pulse1	pulse2	pulse3	diet	id
1	1	112	166	215	1	1
2	1	111	166	225	1	2
3	1	89	132	189	1	3
4	1	95	134	186	2	4

```
#first, reshape data from wide to long
long_data = reshape(data = data01, varying = c("pulse1", "pulse2", "pulse3"), v.names = "pulse",
                    timevar = "time", times = c("time1", "time2", "time3"), direction = "long")
```

	exertype	diet	id	time	pulse
.time1	1	1	1	time1	112
.time1	1	1	2	time1	111
.time1	1	1	3	time1	89
.time1	1	2	4	time1	95
.time1	1	2	5	time1	66
.time1	1	2	6	time1	69
.time1	2	1	7	time1	125
.time1	2	1	8	time1	85

Unconditional Model: Independence Assumption Syntax

```
#empty model
anova_emtpy = lm(pulse ~ 1 + time, data = long_data)
summary(anova_emtpy)

#estimated variance:
summary(anova_emtpy)$sigma^2
```

```
> summary(anova_emtpy)

Call:
lm(formula = pulse ~ 1 + time, data = long_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-58.556 -12.444  -2.306  13.958  51.444 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   87.500      5.311  16.476  < 2e-16 ***
timetime2     46.611      7.510   6.206 9.74e-08 ***
timetime3    102.056      7.510  13.588  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.53 on 51 degrees of freedom
Multiple R-squared:  0.784,	Adjusted R-squared:  0.7755 
F-statistic: 92.55 on 2 and 51 DF,  p-value: < 2.2e-16

> 
> #estimated variance:
> summary(anova_emtpy)$sigma^2
[1] 507.6612
```

Unconditional Model: Independence Assumption

- Parsing through the output of the previous page, we can determine the following for our data:

$$f(\mathbf{y}_i|\mathbf{X}_i) \sim N_p(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V})$$

- Where

$$\widehat{\boldsymbol{\beta}}^T = \begin{bmatrix} 87.5 & 46.6 & 102.1 \end{bmatrix}$$

- And

$$\mathbf{V} = \sigma_e^2 \mathbf{I}_3 = \begin{bmatrix} 507.7 & 0 & 0 \\ 0 & 507.7 & 0 \\ 0 & 0 & 507.7 \end{bmatrix}$$

Repeated Measures Model

- The repeated measures version of the unconditional model changes the structure of the $\mathbf{V}$ matrix:

$$\mathbf{V} = \begin{bmatrix} \sigma_e^2 + \tau & \tau & \tau \\ \tau & \sigma_e^2 + \tau & \tau \\ \tau & \tau & \sigma_e^2 + \tau \end{bmatrix}$$

- The design above is called **compound symmetry**
  - <u>Diagonal elements:</u> variance of outcomes has error variance and compound symmetry parameter
  - <u>Off-diagonal elements:</u> compound symmetry parameter gives covariance of observations

- Because observations come from the same person observations are likely to be correlated
  - **<u>Repeated measures incorporates this correlation through sphericity (a weaker form of compound symmetry)</u>**

Unconditional Model: Repeated Measures CS Assumption

```
#empty model:
RManova_empty = aov(pulse ~ time + Error(id/time), data=long_data)
summary(RManova_empty)
coef(RManova_empty)
```

Impossible to see covariance structure from R output

```
> summary(RManova_empty)

Error: id
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 17  23369    1375

Error: id:time
          Df Sum Sq Mean Sq F value Pr(>F)
time       2  93972   46986   633.5 <2e-16 ***
Residuals 34   2522      74
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> coef(RManova_empty)
(Intercept) :
(Intercept)
   137.0556

id :
numeric(0)

id:time :
timetime2 timetime3
 46.61111 102.05556
```

Comparing Independence vs CS Output

```
> anova(anova_emtpy)
Analysis of Variance Table

Response: pulse
          Df Sum Sq Mean Sq F value    Pr(>F)
time       2  93972   46986  92.554 < 2.2e-16 ***
Residuals 51  25891     508
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> summary(anova_emtpy)

Call:
lm(formula = pulse ~ 1 + time, data = long_data)

Residuals:
    Min      1Q  Median      3Q     Max
-58.556 -12.444  -2.306  13.958  51.444

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   87.500      5.311  16.476  < 2e-16 ***
timetime2     46.611      7.510   6.206 9.74e-08 ***
timetime3    102.056      7.510  13.588  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.53 on 51 degrees of freedom
Multiple R-squared:  0.784,	Adjusted R-squared:  0.7755
F-statistic: 92.55 on 2 and 51 DF,  p-value: < 2.2e-16

>
> #estimated variance:
> summary(anova_emtpy)$sigma^2
[1] 507.6612
```

```
> summary(RManova_empty)

Error: id
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 17  23369    1375

Error: id:time
          Df Sum Sq Mean Sq F value Pr(>F)
time       2  93972   46986   633.5 <2e-16 ***
Residuals 34   2522      74
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> coef(RManova_empty)
(Intercept) :
(Intercept)
   137.0556

id :
numeric(0)

id:time :
timetime2 timetime3
 46.61111 102.05556
```

The Mixed Model Version of RM ANOVA (Part 1)

- We can get the same results from the gls() function:

```
#empty model using mixed model via GLS function:
Mixed_RManova_empty = gls(model = pulse ~ time, method = "REML", data = long_data, correlation=corCompSymm(form=~1|id))
summary(Mixed_RManova_empty)
anova(Mixed_RManova_empty)
getVarCov(Mixed_RManova_empty)
```

```
> summary(Mixed_RManova_empty)
Generalized least squares fit by REML
  Model: pulse ~ time
  Data: long_data
       AIC      BIC    logLik
  432.6616 442.3207 -211.3308

Correlation Structure: Compound symmetry
 Formula: ~1 | id
 Parameter estimate(s):
      Rho
0.8538923

Coefficients:
                Value Std.Error  t-value p-value
(Intercept)  87.50000  5.310687 16.47621       0
timetime2    46.61111  2.870796 16.23630       0
timetime3   102.05556  2.870796 35.54957       0

 Correlation:
          (Intr) timtm2
timetime2 -0.27
timetime3 -0.27   0.50

Standardized residuals:
       Min         Q1        Med         Q3        Max
-2.5988494 -0.5523171 -0.1023266  0.6195075  2.2832396

Residual standard error: 22.53134
Degrees of freedom: 54 total; 51 residual
```

```
> anova(Mixed_RManova_empty)
Denom. DF: 51
            numDF  F-value p-value
(Intercept)     1 737.9024  <.0001
time            2 633.4640  <.0001
```

Comparing RM ANOVA with the LME Model Version

```
> summary(RManova_empty)

Error: id
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 17  23369    1375

Error: id:time
          Df Sum Sq Mean Sq F value Pr(>F)
time       2  93972   46986   633.5 <2e-16 ***
Residuals 34   2522      74
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> coef(RManova_empty)
(Intercept) :
(Intercept)
   137.0556

id :
numeric(0)

id:time :
timetime2 timetime3
 46.61111 102.05556
```

```
> anova(Mixed_RManova_empty)
Denom. DF: 51
            numDF  F-value p-value
(Intercept)     1 737.9024  <.0001
time            2 633.4640  <.0001
```

```
Coefficients:
                Value Std.Error  t-value p-value
(Intercept)  87.50000  5.310687 16.47621       0
timetime2    46.61111  2.870796 16.23630       0
timetime3   102.05556  2.870796 35.54957       0
```

Investigating the V Matrix from the LME Model

- Using the GetVarCov() function for the LME model, we can see the form of the (residual) covariance matrix:

```
> getVarCov(Mixed_RManova_empty)
Marginal variance covariance matrix
       [,1]   [,2]   [,3]
[1,] 507.66 433.49 433.49
[2,] 433.49 507.66 433.49
[3,] 433.49 433.49 507.66
  Standard Deviations: 22.531 22.531 22.531
```

Unconditional Model: Independence Assumption

- Parsing through the output of the previous page, we can determine the following for our data:

$$f(\mathbf{y}_i|\mathbf{X}_i) \sim N_p(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V})$$

- Where

$$\widehat{\boldsymbol{\beta}}^T = [87.5 \quad 46.6 \quad 102.1]$$

- And

$$\mathbf{V} = \begin{bmatrix} \sigma_e^2 + \tau & \tau & \tau \\ \tau & \sigma_e^2 + \tau & \tau \\ \tau & \tau & \sigma_e^2 + \tau \end{bmatrix} = \begin{bmatrix} 74.2 + 433.5 & 433.5 & 433.5 \\ 433.5 & 507.7 & 433.5 \\ 433.5 & 433.5 & 507.7 \end{bmatrix}$$

(RE)ML Allows for Testing Covariance Structures

- With the LME model we can see if the CS covariance structure improved model fit from the independence model:

```
#emtpy independence model with REML
Mixed_anova_empty = gls(model = pulse ~ time, method = "REML", data = long_data)

#comparing independence vs CS
anova(Mixed_anova_empty, Mixed_RManova_empty)
```

- The null hypothesis is that $\tau = 0$

```
> anova(Mixed_anova_empty, Mixed_RManova_empty)
                    Model df      AIC      BIC    logLik   Test  L.Ratio p-value
Mixed_anova_empty       1  4 479.1234 486.8507 -235.5617
Mixed_RManova_empty     2  5 432.6616 442.3207 -211.3308 1 vs 2 48.46176  <.0001
```

- We reject the null hypothesis

Repeated Measures: Full Analysis

- The full repeated measures analysis of our data gives:
  - Between subjects effects
    - Differences between Diet and Exercise Type
  - Within subjects effects
    - Differences between when pulse rate was taken
    - Not a consideration in Multivariate GLM (MANOVA)
  - Interactions between within and between subjects effects
- Within subjects effects can have adjustments to p-values
  - Just in case data do not meet assumption of sphericity
  - SAS and SPSS have adjustments built-in…R does not
- With LME models, we don't need adjustments
  - We have the ability to fit (and test the fit of) the right model!

Full Repeated Measures ANOVA Model

```
#full model:
RManova_full = aov(pulse ~ exertype + diet + exertype*diet + Error(id/time) + (time + time*diet + time*exertype + time*diet*exertype), data=long_data)
summary(RManova_full)
```

```
> summary(RManova_full)

Error: id
              Df Sum Sq Mean Sq F value Pr(>F)
exertype       2   1560     780   0.761 0.4884
diet           1   8791    8791   8.577 0.0126 *
exertype:diet  2    718     359   0.351 0.7113
Residuals     12  12299    1025
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: id:time
                   Df Sum Sq Mean Sq F value Pr(>F)
time                2  93972   46986 703.716 <2e-16 ***
diet:time           2    345     172   2.583 0.0964 .
exertype:time       4     81      20   0.302 0.8740
exertype:diet:time  4    494     123   1.850 0.1523
Residuals          24   1602      67
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Same Analysis with LME Model

```
#RMmanova using Mixed Model via GLS function:
Mixed_RManova_full = gls(model = pulse ~ exertype + diet + exertype*diet +
                     time + time*diet + time*exertype + time*diet*exertype,
                     method = "REML", data = long_data, correlation=corCompSymm(form=~1|id))
summary(Mixed_RManova_full)
anova(Mixed_RManova_full)
getVarCov(Mixed_RManova_full)
```

```
> anova(Mixed_RManova_full)
Denom. DF: 36
                   numDF  F-value p-value
(Intercept)            1 989.6974  <.0001
exertype               2   0.7612  0.4745
diet                   1   8.5775  0.0059
time                   2 703.7157  <.0001
exertype:diet          2   0.3505  0.7067
diet:time              2   2.5830  0.0895
exertype:time          4   0.3016  0.8749
exertype:diet:time     4   1.8495  0.1407
```

The ANOVA Table: Same Results (for F-values)

```
> summary(RManova_full)

Error: id
              Df Sum Sq Mean Sq F value Pr(>F)
exertype       2   1560     780   0.761 0.4884
diet           1   8791    8791   8.577 0.0126 *
exertype:diet  2    718     359   0.351 0.7113
Residuals     12  12299    1025
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Error: id:time
                   Df Sum Sq Mean Sq F value Pr(>F)
time                2  93972   46986 703.716 <2e-16 ***
diet:time           2    345     172   2.583 0.0964 .
exertype:time       4     81      20   0.302 0.8740
exertype:diet:time  4    494     123   1.850 0.1523
Residuals          24   1602      67
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> anova(Mixed_RManova_full)
Denom. DF: 36
                   numDF  F-value p-value
(Intercept)            1 989.6974  <.0001
exertype               2   0.7612  0.4745
diet                   1   8.5775  0.0059
time                   2 703.7157  <.0001
exertype:diet          2   0.3505  0.7067
diet:time              2   2.5830  0.0895
exertype:time          4   0.3016  0.8749
exertype:diet:time     4   1.8495  0.1407
```

- The p-values differ due to LME having multiple methods for determining the denominator DF

The LME Residual Covariance Matrix

- From the Empty model:

```
> getVarCov(Mixed_RManova_empty)
Marginal variance covariance matrix
       [,1]   [,2]   [,3]
[1,] 507.66 433.49 433.49
[2,] 433.49 507.66 433.49
[3,] 433.49 433.49 507.66
  Standard Deviations: 22.531 22.531 22.531
```

- From the Full model:

```
> getVarCov(Mixed_RManova_full)
Marginal variance covariance matrix
       [,1]   [,2]   [,3]
[1,] 386.15 319.38 319.38
[2,] 319.38 386.15 319.38
[3,] 319.38 319.38 386.15
  Standard Deviations: 19.651 19.651 19.651
```

- Note: with CS assumption R^2 is the same for all DVs:
  - = 507.66-386.15/507.66 = .239

- Multivariate version:

```
> #get multivariate R^2:
> full = getVarCov(Mixed_RManova_full); gvar_full = det(matrix(full, nrow=3, ncol=3))
> empty = getVarCov(Mixed_RManova_empty); gvar_empty = det(matrix(empty, nrow=3, ncol=3))
>
> (gvar_empty-gvar_full)/gvar_empty
[1] 0.3958484
```

CLASSICAL MANOVA VS LME MODELS

Multivariate Approaches

- Like repeated measures, multivariate approaches focus on all **dependent variables**, simultaneously
  - Repeated measures analysis

- In multivariate GLM, the effects of interest are:
  - Between subjects effects – Implies differences in mean vectors
    - Differences between Diet and Exercise Type

- Within subjects effects are less commonly inspected
  - But certainly can be

- Key difference between multivariate approach and repeated measures approach comes from assumptions about $\mathbf{V}$ matrix

Multivariate Approach Assumptions

- MANOVA estimates all elements of the **V** matrix:

$$\mathbf{V} = \begin{bmatrix} \sigma^2_{e_1} & \sigma_{e_1 e_2} & \sigma_{e_1 e_3} \\ \sigma_{e_1 e_2} & \sigma^2_{e_2} & \sigma_{e_2 e_3} \\ \sigma_{e_1 e_3} & \sigma_{e_2 e_3} & \sigma^2_{e_3} \end{bmatrix}$$

- Contrast that with the repeated measures version of compound symmetry structure for the **V** matrix:

$$\mathbf{V} = \begin{bmatrix} \sigma^2_e + \tau & \tau & \tau \\ \tau & \sigma^2_e + \tau & \tau \\ \tau & \tau & \sigma^2_e + \tau \end{bmatrix}$$

Multivariate Versus Repeated Measures

- Multivariate:
  - More model parameters; less parsimony
  - Fewer assumptions
  - Better if assumptions of RM are violated (LME models don't need this)

- Repeated measures:
  - Fewer model parameters; more parsimony
  - Strict assumptions
  - More power that multivariate if assumptions are met

Multivariate Test Statistic(s)

- The multivariate approach has a new class of statistics used for testing multivariate hypotheses
  - Rely on summaries of key matrix products
- More than one test statistic is available
  - None are uniformly most powerful (across all sample sizes and types of independent variables)
- Test statistics:
  - Wilks' lambda
  - Pillai's trace
  - Hotelling-Lawley trace
  - Roy's largest root

Forming Multivariate Test Statistics

- As with univariate linear models, multivariate linear models decompose variability into various sources
  - But now decomposition is multivariate

- Univariate decomposition:
  - Sums of squares treatment: $\sum_{g=1}^{k} n_g \left(\bar{y}_{\cdot g} - \bar{y}_{\cdot\cdot}\right)^2$
  - Sums of squares error: $\sum_{i=1}^{n_g} \sum_{g=1}^{k} \left(y_{ig} - \bar{y}_{\cdot g}\right)^2$

- F-ratio was formed by comparing these terms (divided by their degrees of freedom)

Multivariate Decomposition

- Multivariate decomposition is based on mean vectors
  - Each variable is represented

- Sums of squares *and* cross-products for treatment (the H matrix…stands for hypothesis):

$$\mathbf{H} = \sum_{g=1}^{k} n_g \left(\bar{\mathbf{y}}_{\cdot g} - \bar{\mathbf{y}}_{\cdot\cdot}\right)\left(\bar{\mathbf{y}}_{\cdot g} - \bar{\mathbf{y}}_{\cdot\cdot}\right)^T$$

- Sums of squares *and* cross-products for error (the E matrix…stands for error):

$$\mathbf{E} = \sum_{i=1}^{n_g} \sum_{g=1}^{k} \left(\mathbf{y}_{ig} - \bar{\mathbf{y}}_{\cdot \mathrm{g}}\right)\left(\mathbf{y}_{ig} - \bar{\mathbf{y}}_{\cdot g}\right)^T$$

Wilks' Lambda

- Once both matrices are formed, Wilks' lambda can be obtained:

$$\lambda^* = \frac{|\mathbf{E}|}{|\mathbf{H} + \mathbf{E}|}$$

- Equivalently (and shown to try to relate to univariate test statistics), Wilks' lambda can be formed from the eigenvalues (λ) of $\mathbf{E}^{-1}\mathbf{H}$:

$$\lambda^* = \prod_{i=1}^{s} \frac{1}{1 + \lambda_i}$$

- Where $s = \min(p, g - 1)$

THE UNIVERSITY OF KANSAS

Multivariate Empty Model

- Returning the focus away from the test statistics, let's examine the multivariate approach using our empty (unconditional) model

```
#MANOVA for an empty model:
MANOVA_empty = manova(cbind(pulse1, pulse2, pulse3) ~ 1, data=data01)
summary.aov(MANOVA_empty)
```

- The results are pretty sparse (no effects tested):

```
> summary.aov(MANOVA_empty)
 Response pulse1 :
            Df Sum Sq Mean Sq F value Pr(>F)
Residuals   17 4758.5  279.91

 Response pulse2 :
            Df Sum Sq Mean Sq F value Pr(>F)
Residuals   17 8041.8  473.05

 Response pulse3 :
            Df Sum Sq Mean Sq F value Pr(>F)
Residuals   17  13090  770.03
```

LME Model Version of MANOVA: Empty Model

```
#MANOVA empty model using mixed model via GLS function:
Mixed_MANOVA_empty = gls(model = pulse ~ time, method = "REML", data = long_data,
                         correlation=corSymm(form=~1|id), weights=varIdent(form = ~1|time))
summary(Mixed_MANOVA_empty)
anova(Mixed_MANOVA_empty)
getVarCov(Mixed_MANOVA_empty)
```

```
> summary(Mixed_MANOVA_empty)
Generalized least squares fit by REML
  Model: pulse ~ time
  Data: long_data
       AIC      BIC    logLik
  415.0356 432.422 -198.5178

Correlation Structure: General
 Formula: ~1 | id
 Parameter estimate(s):
 Correlation:
  1     2
2 0.917
3 0.852 0.946
Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | time
 Parameter estimates:
   time1    time2    time3
1.000000 1.299993 1.658602

Coefficients:
                Value Std.Error  t-value p-value
(Intercept)  87.50000  3.943432 22.18879       0
timetime2    46.61111  2.184526 21.33694       0
timetime3   102.05556  3.789795 26.92904       0

 Correlation:
          (Intr) timtm2
timetime2 0.346
timetime3 0.430  0.818

Standardized residuals:
       Min         Q1        Med         Q3        Max
-2.1101602 -0.5894138 -0.1125977  0.5650097  2.2414067

Residual standard error: 16.73056
Degrees of freedom: 54 total; 51 residual
```

```
> anova(Mixed_MANOVA_empty)
Denom. DF: 51
            numDF  F-value p-value
(Intercept)     1 137.5151  <.0001
time            2 363.3109  <.0001
```

```
> getVarCov(Mixed_MANOVA_empty)
Marginal variance covariance matrix
       [,1]   [,2]   [,3]
[1,] 279.91 333.53 395.71
[2,] 333.53 473.05 571.23
[3,] 395.71 571.23 770.03
  Standard Deviations: 16.731 21.75 27.749
```

Note: The diagonal elements are equal to the MSEs from the previous slide

THE UNIVERSITY OF KANSAS

Empty Model: Unstructured Covariance Matrix

- Parsing through the output of the previous page, we can determine the following for our data:

$$f(\mathbf{y}_i | \mathbf{X}_i) \sim N_p(\mathbf{X}_i \boldsymbol{\beta}, \mathbf{R})$$

- Where

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} 87.5 & 46.6 & 102.1 \end{bmatrix}$$

- And

$$\mathbf{R} = \left(\frac{1}{N-1}\right)\mathbf{E} = \begin{bmatrix} \sigma_{e_1}^2 & \sigma_{e_1 e_2} & \sigma_{e_1 e_3} \\ \sigma_{e_1 e_2} & \sigma_{e_2}^2 & \sigma_{e_2 e_3} \\ \sigma_{e_1 e_3} & \sigma_{e_2 e_3} & \sigma_{e_3}^2 \end{bmatrix}$$

$$= \begin{bmatrix} 279.9 & 333.5 & 395.7 \\ 333.5 & 473.0 & 571.2 \\ 395.7 & 571.2 & 770.0 \end{bmatrix}$$

Building the Full Model: Classical MANOVA

```
#the full model:
MANOVA_full = manova(cbind(pulse1, pulse2, pulse3) ~ diet + exertype + diet*exertype, data=data01)
summary.manova(MANOVA_full, test="Wilks")
summary.manova(MANOVA_full, test="Pillai")
summary.manova(MANOVA_full, test="Hotelling-Lawley")
summary.manova(MANOVA_full, test="Roy")
```

```
> summary.manova(MANOVA_full, test="Wilks")
              Df   Wilks approx F num Df den Df  Pr(>F)
diet           1 0.54857   2.7430      3     10 0.09885 .
exertype       2 0.60547   0.9505      6     20 0.48224
diet:exertype  2 0.54777   1.1705      6     20 0.36068
Residuals     12
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary.manova(MANOVA_full, test="Pillai")
              Df  Pillai approx F num Df den Df  Pr(>F)
diet           1 0.45143   2.7430      3     10 0.09885 .
exertype       2 0.42936   1.0024      6     22 0.44865
diet:exertype  2 0.47929   1.1557      6     22 0.36471
Residuals     12
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> summary.manova(MANOVA_full, test="Hotelling-Lawley")
              Df Hotelling-Lawley approx F num Df den Df  Pr(>F)
diet           1          0.82291  2.74302      3     10 0.09885 .
exertype       2          0.59408  0.89113      6     18 0.52181
diet:exertype  2          0.77616  1.16424      6     18 0.36766
Residuals     12
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary.manova(MANOVA_full, test="Roy")
              Df     Roy approx F num Df den Df  Pr(>F)
diet           1 0.82291   2.7430      3     10 0.09885 .
exertype       2 0.47228   1.7317      3     11 0.21815
diet:exertype  2 0.70619   2.5894      3     11 0.10580
Residuals     12
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Building the Full Model: LME Model

```
#MANOVA using Mixed Model via GLS function:
#empty model:
Mixed_MANOVA_full = gls(model = pulse ~ exertype + diet + exertype*diet +
                    time + time*diet + time*exertype + time*diet*exertype,
                    method = "REML", data = long_data, correlation=corSymm(form=~1|id),
                    weights=varIdent(form = ~1|time))
summary(Mixed_MANOVA_full)
anova(Mixed_MANOVA_full)
getVarCov(Mixed_MANOVA_full)
```

```
> anova(Mixed_MANOVA_full)
Denom. DF: 36
                   numDF  F-value p-value
(Intercept)            1 268.5689  <.0001
exertype               2   2.2173  0.1236
diet                   1   6.8163  0.0131
time                   2 416.3629  <.0001
exertype:diet          2   1.9365  0.1589
diet:time              2   1.5293  0.2304
exertype:time          4   0.6736  0.6146
exertype:diet:time     4   1.3602  0.2670
```

```
> getVarCov(Mixed_MANOVA_full)
Marginal variance covariance matrix
       [,1]   [,2]   [,3]
[1,] 189.61 244.31 278.64
[2,] 244.31 376.00 435.19
[3,] 278.64 435.19 592.83
  Standard Deviations: 13.77 19.391 24.348
```

THE BENEFITS OF MIXED MODELS OVER CLASSICAL RM ANOVA AND MANOVA

So...Which V Matrix is Right?

Classical methods taught today have no good way of determining which is most appropriate $\mathbf{V}$ matrix

Independence

$$\begin{bmatrix} 507.7 & 0 & 0 \\ 0 & 507.7 & 0 \\ 0 & 0 & 507.7 \end{bmatrix}$$

Regular ANOVA

Compound Symmetry

$$\begin{bmatrix} 507.7 & 433.5 & 433.5 \\ 433.5 & 507.7 & 433.5 \\ 433.5 & 433.5 & 507.7 \end{bmatrix}$$

Repeated Measures ANOVA

Unstructured

$$\begin{bmatrix} 279.9 & 333.5 & 395.7 \\ 333.5 & 473.0 & 571.2 \\ 395.7 & 571.2 & 770.0 \end{bmatrix}$$

Multivariate ANOVA

$\longleftrightarrow$

Restrictive Assumptions

Few Parameters

Relaxed Assumptions

More Parameters

LME Models Can Test for the Best Structure

- We can use Likelihood Ratio Tests to test for the best covariance matrix structure in LME models:

```
> anova(Mixed_RManova_full, Mixed_MANOVA_full)
                   Model df      AIC      BIC    logLik   Test  L.Ratio p-value
Mixed_RManova_full     1 20 345.9564 377.6268 -152.9782
Mixed_MANOVA_full      2 24 336.0586 374.0630 -144.0293 1 vs 2 17.89789  0.0013
>
```

- The LRT above is not rescaled (not available in nlme package: BAD R!)

- We could also come up with a new structure altogether
  - Not at all possible in classical approach

- On next slide: Heterogeneous Compound Symmetry
  - Different covariances per DV
  - Same correlation

New Model: Heterogeneous Compound Symmetry

```
#New Structure: Heterogeneous variances with common correlation (Called CS-H in SAS)
Mixed_CSH_full = gls(model = pulse ~ exertype + diet + exertype*diet +
                     time + time*diet + time*exertype + time*diet*exertype,
                     method = "REML", data = long_data, correlation=corCompSymm(form=~1|id),
                     weights=varIdent(form = ~1|time))
```

```
> getVarCov(Mixed_CSH_full)
Marginal variance covariance matrix
       [,1]   [,2]   [,3]
[1,] 193.32 235.83 303.45
[2,] 235.83 363.56 416.13
[3,] 303.45 416.13 601.89
  Standard Deviations: 13.904 19.067 24.533
```

```
> CSHcorrV
          [,1]      [,2]      [,3]
[1,] 1.0000000 0.8895744 0.8895744
[2,] 0.8895744 1.0000000 0.8895744
[3,] 0.8895744 0.8895744 1.0000000
```

```
> MANOVAcorrV
          [,1]      [,2]      [,3]
[1,] 1.0000000 0.9149714 0.8310816
[2,] 0.9149714 1.0000000 0.9217711
[3,] 0.8310816 0.9217711 1.0000000
```

```
> anova(Mixed_RManova_full, Mixed_CSH_full, Mixed_MANOVA_full)
                   Model df      AIC      BIC    logLik   Test  L.Ratio p-value
Mixed_RManova_full     1 20 345.9564 377.6268 -152.9782
Mixed_CSH_full         2 22 336.0467 370.8841 -146.0233 1 vs 2 13.90974  0.0010
Mixed_MANOVA_full      3 24 336.0586 374.0630 -144.0293 2 vs 3  3.98815  0.1361
>
```

LME MODEL EXTENSIONS: RANDOM COEFFICIENTS

An À la carte Approach to Model Building

Choices for Model for the Means

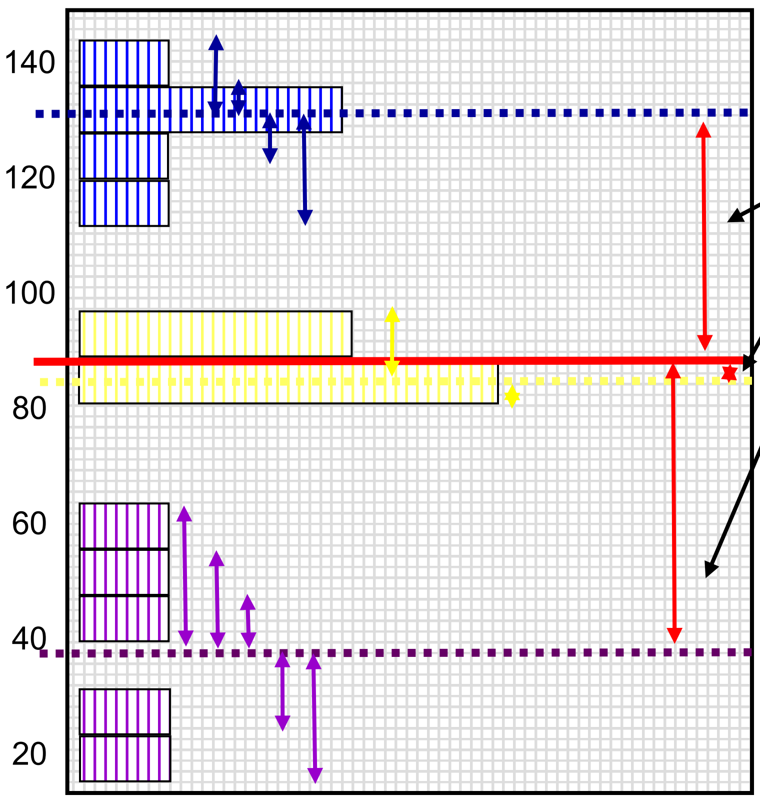
Choices for Model for the Variances	**Empty Means**	**Saturated Means**
e_{ti} only	1a	1b *BP ANOVA*
$U_{0i} + e_{ti}$ Compound Symmetry (CS)	2a	2b *Univ. RM ANOVA*
All variances and covariances (Unstructured; UN)	3a	3b *Multiv. RM ANOVA*

The labels for the models (1a – 3b) correspond to example 2c.

MLM generally begins with 2a, which is used as a baseline model.

THE UNIVERSITY OF KANSAS

Empty Multilevel Model ($U_{0i} + e_{ti}$): New Terminology



Variance of Y → 2 sources:

<u>Level 2 Random Intercept Variance</u> (of U_{0i}):

- → Between-Person Variance (τ_{U0}^2)
- → Difference from GRAND mean
- → INTER-Individual Differences

<u>Level 1 Residual Variance</u> (of e_{ti}):

- → Within-Person Variance (σ_e^2)
- → Difference from OWN mean
- → INTRA-Individual Differences

Empty Multilevel Model

Model for Means; Model for Variances

General Linear Model

$$y_i = \beta_0 + e_i$$

Multilevel Model

L1: $y_{ti} = \beta_{0i} + e_{ti}$

L2: $\beta_{0i} = \gamma_{00} + U_{0i}$

Sample Grand Mean **Intercept**

Individual **Intercept Deviation**

3 Model Parameters

1 Fixed Effect:

γ_{00} → fixed intercept

1 Random Effect (intercept): U_{0i} → person-specific deviation

→ mean=0, variance = τ_{U0}^2

1 Residual Error:

e_{ti} → time-specific deviation

→ mean=0, variance = σ_e^2

Combined equation: $y_{ti} = \gamma_{00} + U_{0i} + e_{ti}$

Empty Multilevel Model: Useful Descriptive Statistic → ICC

IntraClass Correlation (ICC):

$$\text{ICC} = \frac{\text{Intercept Variance}}{\text{Intercept Variance} + \text{Residual Variance}}$$

$$\text{ICC} = \frac{\text{Between Variance}}{\text{Between Variance} + \text{Within Variance}} = \frac{\tau^2_{U_0}}{\tau^2_{U_0} + \sigma^2_e}$$

- ICC = Proportion of variance that is between-persons
- ICC = Average correlation across occasions

THE UNIVERSITY OF KANSAS

Matrices Used in Longitudinal Models for the Variances

- Two matrices of variance components (piles):
  - **G Matrix**: Between-person $\mathbf{U_{0i}}$ variance
  - **R Matrix**: Within-person $\mathbf{e_{ti}}$ variance
  - **G** and **R** combine to make **V**, the total variance of Y

- The most basic +WP model (just $\mathbf{U_{0i}}$ + $\mathbf{e_{ti}}$) can be estimated in two equivalent ways:
  - **"Random Intercept Only Model":** G and R together = V
    - Random intercept only in G
    - "Identity" (SPSS) or "VC" (SAS) R (uncorrelated, homogeneous var)
  - **"Compound Symmetry Model":** Just use R (so R = V)
    - Nothing in G, R has "compound symmetry" form

Random Intercept Model (4 occasions):

$$y_{ti} = \gamma_{00} + U_{0i} + e_{ti}$$

Total Observed Data Matrix is called V Matrix

$$\begin{bmatrix} \sigma^2_{11} & & & \\ \sigma_{21} & \sigma^2_{22} & & \\ \sigma_{31} & \sigma_{32} & \sigma^2_{33} & \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma^2_{44} \end{bmatrix}$$

Level 2, BP Variance

Unstructured **G Matrix**
(RANDOM statement)

$$\begin{array}{ll} & \text{Intercept Variance} \\ \text{Intercept Variance} & \begin{pmatrix} \tau^2_{U_0} \end{pmatrix} \end{array}$$

Level 1, WP Variance

Identity (VC) **R Matrix**
(REPEATED statement)

$$\begin{bmatrix} \begin{matrix} \sigma^2_e & & & \\ 0 & \sigma^2_e & & \\ 0 & 0 & \sigma^2_e & \\ 0 & 0 & 0 & \sigma^2_e \end{matrix} & \\ \begin{matrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{matrix} & \begin{matrix} \sigma^2_e & & & \\ 0 & \sigma^2_e & & \\ 0 & 0 & \sigma^2_e & \\ 0 & 0 & 0 & \sigma^2_e \end{matrix} \end{bmatrix}$$

Each person gets same 4 x 4 R matrix → **equal variances and 0 covariances** across persons & time

Equivalent Variance Models

- **"Random Intercept Only Model": G and R = V**
  - Allows quantifiable estimates of BP vs. WP variation (separate estimates) and constant correlation between persons over time

Get "Var(e_{ti})" and "Var(U_{0i})"

G matrix $\begin{pmatrix} \tau^2_{U_0} \end{pmatrix}$ "+" **R matrix** $\begin{bmatrix} \sigma^2_e & & & \\ 0 & \sigma^2_e & & \\ 0 & 0 & \sigma^2_e & \\ 0 & 0 & 0 & \sigma^2_e \end{bmatrix}$ $\rightarrow$ **Total V matrix** $\begin{bmatrix} \sigma^2_e+\tau^2_{u_0} & & & \\ \tau^2_{u_0} & \sigma^2_e+\tau^2_{u_0} & & \\ \tau^2_{u_0} & \tau^2_{u_0} & \sigma^2_e+\tau^2_{u_0} & \\ \tau^2_{u_0} & \tau^2_{u_0} & \tau^2_{u_0} & \sigma^2_e+\tau^2_{u_0} \end{bmatrix}$

- **"Compound Symmetry Model (CS)": R = V**
  - BP and WP variance combined into one estimate instead, but still allows constant correlation between persons over time

Get "Var(e_{ti})" and "CS" → CS = Var(U_{0i})

R matrix $\begin{bmatrix} \sigma^2_T & & & \\ CS & \sigma^2_T & & \\ CS & CS & \sigma^2_T & \\ CS & CS & CS & \sigma^2_T \end{bmatrix}$ = **Total V matrix** $\begin{bmatrix} \sigma^2_e+\tau^2_{u_0} & & & \\ \tau^2_{u_0} & \sigma^2_e+\tau^2_{u_0} & & \\ \tau^2_{u_0} & \tau^2_{u_0} & \sigma^2_e+\tau^2_{u_0} & \\ \tau^2_{u_0} & \tau^2_{u_0} & \tau^2_{u_0} & \sigma^2_e+\tau^2_{u_0} \end{bmatrix}$

ANALYSIS ADVICE

Advice About Picking An Analysis Method

- Use REPEATED MEASURES if:
  - You have complete data (nothing is missing)
  - You have a small sample size
  - Your data meet sphericity assumption

- Use MULTIVARIATE if:
  - You have complete data (nothing is missing)
  - You have a large sample size
  - Your data do not meet sphericity assumption

- Note: both methods, although classical, are still useful
  - However: we will learn a modern approach to their estimation
    - Allows for more broad analysis types and missing data
  - Problem: cannot tell which $\mathbf{R}$ matrix is appropriate

THE UNIVERSITY OF KANSAS

When Not To Use Either Method

- Do NOT use this version of REPEATED MEASURES or MULTIVARIATE if:
  - You have independent variables you wish to study that vary with each dependent variable (example: time)
    - Use multilevel approach
  - You have missing data
    - We will use this to demonstrate imputation methods
  - You have categorical data
    - Use link-functions
  - You have data that come from psychometric measures
    - Use structural equation modeling
  - You would like to determine which $\mathbf{R}$ matrix to use

CONCLUDING REMARKS

Wrapping Up

- Many methods for repeated measures exist
  - Univariate
  - Repeated measures
  - Multivariate

- What you have learned to this point in this class can
  - Do repeated measures more generally than classical methods
  - Be more flexible with missing data
  - Accommodate less restrictive assumptions