Introduction to Modeling Multivariate Outcomes with Maximum Likelihood Part 1: Conditionally Multivariate Normal Outcomes

PSYC 943 (930): Fundamentals of Multivariate Modeling

Lecture 12: October 5, 2012

Today's Class

- General linear models in matrices for…
  - One (univariate) conditionally normal outcome
  - Estimated using maximum likelihood in PROC MIXED
- Expanding linear models from univariate outcomes to multivariate outcomes:
  - Multiple variable analyses, simultaneously
  - All outcomes are then assumed to be conditionally multivariate normally distributed
- Models for covariances

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

Original Data: Wide Format

- The data:

	Exercise Type	Pulse After Warmup	Pulse After Jogging	Pulse After Running	Diet Type	personID
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
5	1	66	109	150	2	5
6	1	69	119	177	2	6
7	2	125	177	241	1	7
8	2	85	117	186	1	8
9	2	97	137	185	1	9
10	2	93	151	217	2	10
11	2	77	122	178	2	11
12	2	78	119	173	2	12
13	3	81	134	205	1	13
14	3	88	133	180	1	14
15	3	88	157	224	1	15
16	3	58	99	131	2	16
17	3	85	132	186	2	17
18	3	78	110	164	2	18

$\bar{x}$
87.5
134.11111
189.55556

S		
264.36111	315	373.72222
315	446.76543	539.49383
373.72222	539.49383	727.24691

Comparing Univariate and Multivariate Normal Distributions

From: Wednesday's Class

- The univariate normal distribution:

$$f(x_p) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

- The univariate normal, rewritten with a little algebra:

$$f(x_p) = \frac{1}{(2\pi)^{\frac{1}{2}}|\sigma^2|^{\frac{1}{2}}} \exp\left[-\frac{(x-\mu)(\sigma_e^2)^{-1}(x-\mu)}{2}\right]$$

- The multivariate normal distribution

$$f(\mathbf{x}_p) = \frac{1}{(2\pi)^{\frac{V}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{\left(\mathbf{x}_p^T - \boldsymbol{\mu}\right)^T \mathbf{\Sigma}^{-1}\left(\mathbf{x}_p^T - \boldsymbol{\mu}\right)}{2}\right]$$

  - When $V = 1$ (one variable), the MVN is a univariate normal distribution

LINEAR MODELS WITH MATRICES

General Linear Models in Matrices

- Matrix expression of the GLM is important in that many descriptions of multivariate statistical models use matrix form
  - This is the starting point for learning the "language of multivariate"

- In this section, we will use an empty model with a single outcome
  - Pulse 3: Pulse after running

Linear Models with Matrices

- The basic linear model for observation *p* (of *N*), as modeled by *k* predictor variables (some of which may be interactions):

$$y_p = \beta_0 + \beta_1 X_{p1} + \cdots + \beta_k X_{pk} + e_p$$

- The equation above, for a single univariate outcome y_p can be expressed more compactly by a set of matrices:

$$y_p = \boldsymbol{x}_p \boldsymbol{\beta} + e_p$$

- y_p is of size (1 x 1) – a scalar (univariate/single outcome)
- $\boldsymbol{x}_p$ is of size (1 x (1 + k)) – the 1 before the + k is for the intercept
- $\boldsymbol{\beta}$ is of size ((1+k) x 1)
- e_p is of size (N x 1) – one outcome means one error per person *p*

Unpacking the Equation

$$\underset{(1\ x\ 1)}{\underset{y_p}{\begin{bmatrix} y_p \end{bmatrix}}} = \underset{(1\ x\ (1+k))}{\underset{\boldsymbol{x}_p}{\begin{bmatrix} 1 & X_{p1} & \cdots & X_{pk} \end{bmatrix}}} \underset{((1+k) x\ 1)}{\underset{\boldsymbol{\beta}}{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}}} + \underset{(1\ x\ 1)}{\underset{e_p}{\begin{bmatrix} e_p \end{bmatrix}}}$$

For any person p:

$$y_p = \beta_0 + \beta_1 X_{p1} + \cdots + \beta_k X_{pk} + e_p$$

Assumed Distributions

- The conditional distribution of y_p has a normal distribution:
  - Mean is the predicted value of y_p (conditional mean)
  - Error variance is the variance of y_p (conditional variance)

$$f\left(y_p \middle| \boldsymbol{x}_p\right) \sim N_1\left(\boldsymbol{x_p}\boldsymbol{\beta}, \sigma_e^2\right)$$

- Because we have only one **dependent variable** we have a univariate normal distribution
  - Mean is determined by (**model for the mean**):
    - **Independent variables**
    - **Linear model coefficients in $\boldsymbol{\beta}$**
  - Variance is determined **only** by σ_e^2 (**model for the variance**)

- This is why checking only the **dependent variable** for normality isn't a good idea
  - Conditional distribution of Y given **X** is normal
  - No assumptions about **X**

The Normal Distribution as a Likelihood Function

- How ML estimation works with conditionally normal outcomes in GLMs is that each person contributes a portion to the total sample log likelihood:

- First, we find the (not-log) likelihood of a single observation

$$L(\sigma_e^2) = \frac{1}{(2\pi)^{\frac{1}{2}}|\sigma_e^2|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(y_p - \hat{y}_p\right)(\sigma_e^2)^{-1}\left(y_p - \hat{y}_p\right)\right)$$

- From that we get the log-likelihood for that same single observation

$$\log L(\sigma_e^2) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log|\sigma_e^2| - \frac{1}{2}\left(y_p - \hat{y}_p\right)(\sigma_e^2)^{-1}\left(y_p - \hat{y}_p\right)$$

- $\hat{y}_p = \boldsymbol{x}_p\boldsymbol{\beta}$, is the **conditional mean** of y_p (model for the means)
- σ_e^2 is the error variance (or the residual variance), the conditional variance of y_p (the model for the variances)

How PROC MIXED Finds Estimates

- For a given value of σ_e^2, there is an equation that provides the fixed effects (model for the means) in $\boldsymbol{\beta}$

$$\boldsymbol{\beta} = (\mathbf{X}^T(\sigma_e^2)^{-1}\mathbf{X})^{-1}\mathbf{X}^T(\sigma_e^2)^{-1}\mathbf{y}$$

- $\mathbf{X}$ is a matrix for all N people with all k predictors (size $N\ x\ k$)
- $\boldsymbol{y}$ is a column vector with all persons outcomes (size $N\ x\ 1$)
- σ_e^2 is the value of the error variance that is currently being evaluated

- For each iteration, PROC MIXED
  1. Finds σ_e^2, then uses it to find $\boldsymbol{\beta}$
  2. Then uses $\boldsymbol{\beta}$ to find $\hat{y}_p$ for all people
  3. Then evaluates the log likelihood

Empty Model in SAS For Pulse After Running

- Syntax:

```
*GLM WITH ML FOR ONE DV (CAN USE WIDE DATA AS ONLY ONE DV IS PRESENT):;
PROC MIXED DATA=WORK.DIETWIDE METHOD=ML COVTEST NOPROFILE ITDETAILS IC NAMELEN=50;
MODEL PULSE3 = /S;
REPEATED / R RCORR;
RUN;
```

  - Although we only have one outcome, the REPEATED line is used to demonstrate how SAS handles the error variances
  - We will shortly use the REPEATED line when we have multiple outcomes

- Output:

CovP1: σ_e^2 for that iteration

Iteration History

CovP1	Iteration	Evaluations	-2 Log Like	Criterion
1.0000	0	1	169.68857616	
727.25	1	1	169.68857616	0.00000000

Convergence criteria met.

More SAS Output

- Parsing the relevant SAS output gives us:

Covariance Parameter Estimates				
Cov Parm	Estimate	Standard Error	Z Value	Pr > Z
Residual	727.25	242.42	3.00	0.0013

  - $\sigma_e^2 = 727.25$
    - As this is the empty model, this is equal to the variance of Y

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	189.56	6.3563	17	29.82	<.0001

  - $\beta_0 = 189.56$
    - As this is the empty model, this is equal to the mean of Y

Putting Output Into Matrices and Likelihoods

- For this analysis, the matrix of all predictors for all people:

$$\mathbf{X} = \mathbf{1}_{(18\,x\,1)}$$

- So, the intercept came from:

$$\boldsymbol{\beta} = (\mathbf{X}^T(\sigma_e^2)^{-1}\mathbf{X})^{-1}\mathbf{X}^T(\sigma_e^2)^{-1}\mathbf{y}$$

$$= \left(\mathbf{1}_{(1\,x\,18)}^T(\sigma_e^2)^{-1}\mathbf{1}_{(18\,x\,1)}\right)^{-1}\mathbf{1}_{(1\,x\,18)}^T(\sigma_e^2)^{-1}\boldsymbol{y}_{(18\,x\,1)}$$

$$= \left(\frac{N}{\sigma_e^2}\right)^{-1}\sum_{p=1}^{18}\frac{y_p}{\sigma_e^2} = \frac{\sigma_e^2}{N}\left(\frac{1}{\sigma_e^2}\right)\sum_{p=1}^{18} y_p = \bar{y}$$

- And…the log likelihood for a person would be:

$$\log L(\sigma_e^2) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log|\sigma_e^2| - \frac{1}{2}\left(y_p - \hat{y}_p\right)(\sigma_e^2)^{-1}\left(y_p - \hat{y}_p\right)$$

$$= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(727.25)$$

$$-\frac{1}{2}\left(y_p - 189.56\right)(727.25)^{-1}\left(y_p - 189.56\right)$$

MULTIVARIATE MODELS

From Univariate to Multivariate

- The first set of slides covered how linear models work when we have a conditional univariate normal outcome
  - In the study, however, there were three outcomes

- We wish to model all three outcomes simultaneously
  - Simultaneous modeling allows for:
    - Determining differences **across** outcomes in addition to differences within outcomes (i.e., as created by predictors)

    - Providing a mechanism to model and simultaneously test:
      - Within subjects factors (pulse rate across intensity levels)
      - Between subjects factors (diet, exercise)
      - Interactions of within and between subjects factors

- Our mechanism for studying the multivariate relationships will be to treat all three outcomes as being part of a (eventually conditional) **multivariate normal distribution**

Multivariate Setup for Data: Stacked (Long) Format

personID	dEXERCISE_ASC	dEXERCISE_R	dEXERCISE_WT	dDIET_M	dDIET_V	intensity	pulse
1	1	0	0	1	0	Pulse: Warm Up	112
1	1	0	0	1	0	Pulse: Jogging	166
1	1	0	0	1	0	Pulse: Running	215
2	1	0	0	1	0	Pulse: Warm Up	111
2	1	0	0	1	0	Pulse: Jogging	166
2	1	0	0	1	0	Pulse: Running	22[illegible]
3	1	0	0	1	0	Pulse: Warm Up	89
3	1	0	0	1	0	Pulse: Jogging	132
3	1	0	0	1	0	Pulse: Running	189
4	1	0	0	0	1	Pulse: Warm Up	95
4	1	0	0	0	1	Pulse: Jogging	134
4	1	0	0	0	1	Pulse: Running	186
5	1	0	0	0	1	Pulse: Warm Up	66
5	1	0	0	0	1	Pulse: Jogging	109
5	1	0	0	0	1	Pulse: Running	150
6	1	0	0	0	1	Pulse: Warm Up	69
6	1	0	0	0	1	Pulse: Jogging	119
6	1	0	0	0	1	Pulse: Running	177
7	0	1	0	1	0	Pulse: Warm Up	125
7	0	1	0	1	0	Pulse: Jogging	177
7	0	1	0	1	0	Pulse: Running	241
8	0	1	0	1	0	Pulse: Warm Up	85
8	0	1	0	1	0	Pulse: Jogging	117
8	0	1	0	1	0	Pulse: Running	186
9	0	1	0	1	0	Pulse: Warm Up	97
9	0	1	0	1	0	Pulse: Jogging	137
9	0	1	0	1	0	Pulse: Running	185
10	0	1	0	0	1	Pu	
10	0	1	0	0	1	Pu	
10	0	1	0	0	1	Pu	
11	0	1	0	0	1	Pu	
11	0	1	0	0	1	Pu	
11	0	1	0	0	1	Pu	
12	0	1	0	0	1	Pu	
12	0	1	0	0	1	Pu	
12	0	1	0	0	1	Pu	
13	0	0	1	1	0	Pu	
13	0	0	1	1	0	Pu	
13	0	0	1	1	0	Pu	
14	0	0	1	1	0	Pu	
14	0	0	1	1	0	Pu	
14	0	0	1	1	0	Pu	
15	0	0	1	1	0	Pu	
15	0	0	1	1	0	Pu	
15	0	0	1	1	0	Pu	
16	0	0	1	0	1	Pu	
16	0	0	1	0	1	Pu	
16	0	0	1	0	1	Pu	
17	0	0	1	0	1	Pu	
17	0	0	1	0	1	Pu	
17	0	0	1	0	1	Pu	
18	0	0	1	0	1	Pu	
18	0	0	1	0	1	Pu	
18	0	0	1	0	1	Pu	

Data for Person #1

Intensity: Variable that denotes which pulse observation is given on that row of data

```
*CONVERTING DATA TO STACKED FORM FOR PROC MIXED;
DATA WORK.dietstack;
    SET WORK.dietwide;

    FORMAT intensity intensities.; *ADDING A FORMAT STATEMENT FOR INTENSITY VARIABLE;

    *FIRST OUTCOME: PULSE 1 (AFTER WARM UP);
    pulse = pulse1;
    intensity = 1;
    dINTENSITY_W = 1; dINTENSITY_J = 0; dINTENSITY_R = 0; *DUMMY CODED VARIABLES FOR ANALYSIS;
    OUTPUT; *OUTPUT MAKES THE LINE OF DATA GET WRITTEN TO THE NEW DATA SET;

    *SECOND OUTCOME: PULSE 2 (AFTER JOGGING);
    pulse = pulse2;
    intensity = 2;
    dINTENSITY_W = 0; dINTENSITY_J = 1; dINTENSITY_R = 0;
    OUTPUT;

    *THIRD OUTCOME: PULSE 3 (AFTER RUNNING);
    pulse = pulse3;
    intensity = 3;
    dINTENSITY_W = 0; dINTENSITY_J = 0; dINTENSITY_R = 1;
    OUTPUT;

RUN;
```

Why Stacked Data?

- Stacked data seem a bit counter-intuitive if you are used to repeated measures types of experiments
  - Most repeated measures analysis programs take wide-format data

- In short, stacked data allow for a more concise method of matching IVs to DVs, making it easy to:
  - Specify if some IVs are different across observations (important in longitudinal research)

  - Keep more data in a maximum likelihood-based analysis if one or more outcomes are missing (see lecture on missing data later in October)
    - HINT: the MVN for a person uses a smaller covariance matrix
    - Use the rows you observe

Multivariate Empty Model

- What a multivariate empty model will give us is very similar to the univariate empty model:
  - The mean for each variable (we can think of this as a mean vector)
    - Three means in our analysis – this will equal our mean vector for this analysis
    - Model for the means now is for a mean vector
  - An estimate of the variance for each variable
    - Three variances in our analysis
  - An estimate of the covariance for each pair variables
    - Three covariances in our analysis – all of these will be equal to our covariance matrix for this analysis
    - The model for the variance is now a model for the covariance matrix
- The trick, in syntax, is to figure out how to get access to all parts
- The trick, in multivariate modeling, is to get an appropriate* model for the covariance matrix so you can believe your model for the means
  - *Appropriate = best fitting and most parsimonious

SAS Syntax for Multivariate Empty Model

```
TITLE "EMPTY MULTIVARIATE MODEL, VC ERROR: (PREDICTORS ARE INDICATORS OF WHICH VARIABLE)";
PROC MIXED DATA=WORK.dietstack METHOD=ML COVTEST NOPROFILE ITDETAILS IC NAMELEN=50;
MODEL pulse = dINTENSITY_W dINTENSITY_R / S DDFM=KENWARDROGER;
REPEATED / SUBJECT=personID TYPE=VC R RCORR;
RUN;
```

- The MODEL line – where the model for the means goes:
  - Pulse still shows up to the left of the equals sign (because pulse is one column now)
  - To the right of the equals sign we now need predictors that will allow us to get a mean estimate for each variable
    - Without predictors here, we would only get one term (an intercept)

$$y_p = \beta_0 + \beta_1 \boldsymbol{dWARMUP}_p + \beta_2 \boldsymbol{dRUNNING}_p + e_p$$

- With your knowledge of linear models, what does:

$\beta_0 =?$

$\beta_1 =?$

$\beta_2 =?$

SAS Syntax for Multivariate Empty Model

```
TITLE "EMPTY MULTIVARIATE MODEL, VC ERROR: (PREDICTORS ARE INDICATORS OF WHICH VARIABLE)";
PROC MIXED DATA=WORK.dietstack METHOD=ML COVTEST NOPROFILE ITDETAILS IC NAMELEN=50;
MODEL pulse = dINTENSITY_W dINTENSITY_R / S DDFM=KENWARDROGER;
REPEATED / SUBJECT=personID TYPE=VC R RCORR;
RUN;
```

- The REPEATED line – provides access to the covariance matrix
  - SUBJECT = personID: Indicates that observations with the same personID are all from the same subject/person (and as such get put into a single multivariate normal distribution)
  - TYPE = VC: The type line gives access to the model for the covariance matrix
    - VC stands for variance components
      - The default, estimates $\sigma_e^2\mathbf{I}$ (or one single residual variance that is shared/the same for each outcome)
      - Does not estimate any residual covariances between outcomes: assumes residuals are independent
    - Many types exist in SAS (http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/statug_mixed_sect020.htm#statug.mixed.repeatedstmt_type) – to be discussed shortly
  - R: SAS' notation for the residual covariance matrix (the letter prints the matrix)
  - RCORR: the correlation matrix version of the R covariance matrix (the word prints the matrix)

Helpful SAS Output Information

- SUBJECTS – should equal your sample size
- Covariance parameters – number of parameters estimated for the covariance matrix (1 = our variance)
- Max Obs Per Subject – should equal your max per subject
- If any of these are off, the model is specified incorrectly in syntax

Dimensions	
Covariance Parameters	1
Columns in X	3
Columns in Z	0
Subjects	18
Max Obs Per Subject	3

Number of Observations	
Number of Observations Read	54
Number of Observations Used	54
Number of Observations Not Used	0

Multivariate Output from PROC MIXED: Fixed Effects

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	134.11	5.1611	54	25.99	<.0001
dINTENSITY_W	-46.6111	7.2988	54	-6.39	<.0001
dINTENSITY_R	55.4444	7.2988	54	7.60	<.0001

- Putting these terms into the matrices from before:

$$\boldsymbol{y}_p = \begin{bmatrix} Pulse1_p \\ Pulse2_p \\ Pulse3_p \end{bmatrix}; \mathbf{X}_p = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix}; \boldsymbol{\beta} = \begin{bmatrix} 134.11 \\ -46.61 \\ 55.44 \end{bmatrix}$$

- Therefore, for any given observation, the predicted (mean vector):

$$\hat{\boldsymbol{y}}_p = \mathbf{X}_p\boldsymbol{\beta} = \begin{bmatrix} 134.11 - 46.61 \\ 134.11 \\ 134.11 + 55.44 \end{bmatrix} = \begin{bmatrix} 87.50 \\ 134.11 \\ 189.56 \end{bmatrix}$$

- These are the means for each outcome from the means vector

Multivariate Output from PROC MIXED: Variances

Covariance Parameter Estimates					
Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > Z
Residual	personID	479.46	92.2717	5.20	<.0001

Estimated R Matrix for Subject 1			
Row	Col1	Col2	Col3
1	479.46		
2		479.46	
3			479.46

Estimated R Correlation Matrix for Subject 1			
Row	Col1	Col2	Col3
1	1.0000		
2		1.0000	
3			1.0000

- The **R** matrix here has one estimated parameter: $\sigma_e^2 = 479.46$
  - Note: look at how large the standard error is – variances are very hard to estimate (and covariances ever harder)…large samples needed
- So, here we find that

$$\mathbf{R} = \sigma_e^2 \mathbf{I}_{(3\ x\ 3)} = 479.46 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 479.46 & 0 & 0 \\ 0 & 479.46 & 0 \\ 0 & 0 & 479.46 \end{bmatrix}$$

Putting Output Into Distributional Terms

- For multivariate data today we are assuming that the multivariate distribution of all outcomes is multivariate normal, conditional on the IVs (although we don't always have to assume conditional MVN)

$$f\left(\boldsymbol{y}_p \middle| \mathbf{X}_p\right) \sim N_V\left(\mathbf{X}_p\boldsymbol{\beta}, \mathbf{V}_p\right)$$

Where:

- $\boldsymbol{y}_p$ is a $V\ x\ 1$ vector of outcomes for person p
- $\mathbf{X}_p$ is a $V\ x\ (k+1)$ matrix of k predictors for person p
- $\boldsymbol{\beta}$ is a $(k+1)x\ 1$ vector of fixed effects
- $\mathbf{X}_p\boldsymbol{\beta} = \widehat{\boldsymbol{y}}_p$ is the predicted conditional mean vector of $\boldsymbol{y}_p$
- $\mathbf{V}_p$ is the residual covariance matrix (SAS notation) for person p

- Today (and for most of this class), we will say $\mathbf{V}_p = \mathbf{R}$ for all people
  - More complicated models bring about more terms into how $\mathbf{V}_p$ is formed
  - In 2-level multilevel models and structural equation models

$$\mathbf{V}_p = \mathbf{Z}_p\mathbf{G}\mathbf{Z}_p^T + \mathbf{R}_p$$

Visualizing the Log-Likelihood for Our Example

- From our example we found:

$$\boldsymbol{\beta} = \begin{bmatrix} 134.11 \\ -46.61 \\ 55.44 \end{bmatrix}; \mathbf{R} = \begin{bmatrix} 479.46 & 0 & 0 \\ 0 & 479.46 & 0 \\ 0 & 0 & 479.46 \end{bmatrix}; \hat{\boldsymbol{y}}_p = \begin{bmatrix} 87.50 \\ 134.11 \\ 189.56 \end{bmatrix}$$

- The conditional MVN for a person is:

$$f\left(\boldsymbol{y}_p | \mathbf{X}_p\right) = \frac{1}{(2\pi)^{\frac{3}{2}} |\mathbf{R}|^{\frac{1}{2}}} \exp\left[-\frac{\left(\boldsymbol{y}_p - \hat{\boldsymbol{y}}_p\right)^T \mathbf{R}^{-1} \left(\boldsymbol{y}_p - \hat{\boldsymbol{y}}_p\right)}{2}\right]$$

- And…the log likelihood for a person would be:

$$\log L(\mathbf{R}) = -\frac{3}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{R}| - \frac{1}{2}\left(\boldsymbol{y}_p - \hat{\boldsymbol{y}}_p\right)^T \mathbf{R}^{-1} \left(\boldsymbol{y}_p - \hat{\boldsymbol{y}}_p\right)$$

$$= -\frac{3}{2}\log(2\pi) - \frac{1}{2}\log(110{,}219{,}172)$$

$$-\frac{1}{2}\left(\begin{bmatrix} Pulse1_p \\ Pulse2_p \\ Pulse3_p \end{bmatrix} - \begin{bmatrix} 87.50 \\ 134.11 \\ 189.56 \end{bmatrix}\right)^T \begin{bmatrix} 0.002 & 0 & 0 \\ 0 & 0.002 & 0 \\ 0 & 0 & 0.002 \end{bmatrix} \left(\begin{bmatrix} Pulse1_p \\ Pulse2_p \\ Pulse3_p \end{bmatrix} - \begin{bmatrix} 87.50 \\ 134.11 \\ 189.56 \end{bmatrix}\right)$$

Inferences from the Mean Vector (Model for the Means)

- We can now turn our attention to what the model for the means is giving us in terms of within –subjects tests
  - To do so, we'll rely on our friend, the ESTIMATE statement

```
PROC MIXED DATA=WORK.dietstack METHOD=ML COVTEST NOPROFILE ITDETAILS IC NAMELEN=50;
MODEL pulse = dINTENSITY_W dINTENSITY_R / S DDFM=KENWARDROGER;
REPEATED / SUBJECT=personID TYPE=UN R RCORR;
ESTIMATE 'Mean Pulse for Warm Up' intercept 1 dINTENSITY_W 1;
ESTIMATE 'Mean Pulse for Jogging' intercept 1;
ESTIMATE 'Mean Pulse for Running' intercept 1 dINTENSITY_R 1;
ESTIMATE 'Warm Up/Jogging Difference' dINTENSITY_W 1;
ESTIMATE 'Warm Up/Running Difference' dINTENSITY_W 1 dINTENSITY_R -1;
ESTIMATE 'Jogging/Running Difference' dINTENSITY_R -1;
CONTRAST 'Test for Overall Within Subject Difference' dINTENSITY_W 1, dINTENSITY_R 1;
RUN;
```

  - Also note the use of the CONTRAST statement
    - A CONTRAST statement is an estimate statement with a multiple degree of freedom test statistic (an F-test)
  - We will use the ESTIMATE statements to do post-hoc differences between within-subjects variables
  - We will use the CONTRAST statement to do an omnibus test of the within-subjects factor

Inferences from the Mean Vector (Model for the Means)

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	134.11	4.9820	18	26.92	<.0001
dINTENSITY_W	-46.6111	2.1230	18	-21.96	<.0001
dINTENSITY_R	55.4444	2.2976	18	24.13	<.0001

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
dINTENSITY_W	1	18	482.05	<.0001
dINTENSITY_R	1	18	582.31	<.0001

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Mean Pulse for Warm Up	87.5000	3.8323	18	22.83	<.0001
Mean Pulse for Jogging	134.11	4.9820	18	26.92	<.0001
Mean Pulse for Running	189.56	6.3563	18	29.82	<.0001
Warm Up/Jogging Difference	-46.6111	2.1230	18	-21.96	<.0001
Warm Up/Running Difference	-102.06	3.6830	18	-27.71	<.0001
Jogging/Running Difference	-55.4444	2.2976	18	-24.13	<.0001

Contrasts

Label	Num DF	Den DF	F Value	Pr > F
Test for Overall Within Subject Difference	2	17	363.31	<.0001

MODELS FOR (THE VARIANCES) COVARIANCE MATRICES

Modeling Covariances

- The estimated covariance matrix in our example analysis was

$$\mathbf{R} = \begin{bmatrix} 479.46 & 0 & 0 \\ 0 & 479.46 & 0 \\ 0 & 0 & 479.46 \end{bmatrix}$$

- From the beginning of class, however, we found the sample covariance matrix to be:

$$\mathbf{S} = \begin{bmatrix} 264.36 & 315.00 & 373.72 \\ 315.00 & 446.75 & 536.49 \\ 373.72 & 539.49 & 727.25 \end{bmatrix}$$

- Deciding on the right model for the covariance matrix is a balance between power and model fit
  - More parameters = (possibly) better fit + less statistical power
- NOTE: Model fit $\neq$ Effect size
  - We are not explaining anything by finding a good fitting model
  - Model fit is necessary, but not sufficient

A (Possibly) Better Model for the Covariance Matrix

```
TITLE "EMPTY MULTIVARIATE MODEL, FULL UNSTRUCTURED ERROR VARIANCE MODEL: (PREDICTORS ARE INDICATORS OF WHICH VARIABLE)
PROC MIXED DATA=WORK.dietstack METHOD=ML COVTEST NOPROFILE ITDETAILS IC NAMELEN=50;
MODEL pulse = dINTENSITY_W dINTENSITY_R / S DDFM=KENWARDROGER;
REPEATED / SUBJECT=personID TYPE=UN R RCORR;
RUN;
```

- REPEATED line: change TYPE = VC (Variance Components) to TYPE = UN (Unstructured)

- An UNSTRUCTURED covariance matrix is one where every term is a model parameter and is estimated:

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

- As this is an empty model, what would you expect the estimates to be?

The Unstructured Model Estimates: Covariance Parameters

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
UN(1,1)	personID	264.36	88.1204	3.00	0.0013
UN(2,1)	personID	315.00	109.88	2.87	0.0041
UN(2,2)	personID	446.77	148.92	3.00	0.0013
UN(3,1)	personID	373.72	135.79	2.75	0.0059
UN(3,2)	personID	539.49	184.99	2.92	0.0035
UN(3,3)	personID	727.25	242.42	3.00	0.0013

Estimated R Matrix for Subject 1

Row	Col1	Col2	Col3
1	264.36	315.00	373.72
2	315.00	446.77	539.49
3	373.72	539.49	727.25

Estimated R Correlation Matrix for Subject 1

Row	Col1	Col2	Col3
1	1.0000	0.9166	0.8523
2	0.9166	1.0000	0.9465
3	0.8523	0.9465	1.0000

Comparing Covariances

- The unstructured model provided a new estimated $\mathbf{R}$ covariance matrix:

$$\mathbf{R}_{UN} = \begin{bmatrix} 264.36 & 315.00 & 373.72 \\ 315.00 & 446.75 & 536.49 \\ 373.72 & 539.49 & 727.25 \end{bmatrix}$$

- The estimated covariance matrix in our example analysis was

$$\mathbf{R}_{VC} = \begin{bmatrix} 479.46 & 0 & 0 \\ 0 & 479.46 & 0 \\ 0 & 0 & 479.46 \end{bmatrix}$$

- From the beginning of class, however, we found the sample covariance matrix to be:

$$\mathbf{S} = \begin{bmatrix} 264.36 & 315.00 & 373.72 \\ 315.00 & 446.75 & 536.49 \\ 373.72 & 539.49 & 727.25 \end{bmatrix}$$

- So, which model is correct: VC or UN?
  - Good news: VC is nested within UN so we can use a likelihood ratio test

Model Comparison

- We will compare the fit of the VC model to the UN model using a likelihood ratio test

$H_0: \mathbf{R} = \sigma_e^2 \mathbf{I}$ (3 fixed effects + 1 variance = 4 parameters)

$$H_A: \mathbf{R} = \begin{bmatrix} \sigma_{e_1}^2 & \sigma_{e_1,e_2} & \sigma_{e_1,e_3} \\ \sigma_{e_1,e_2} & \sigma_{e_2}^2 & \sigma_{e_2,e_3} \\ \sigma_{e_1,e_3} & \sigma_{e_2,e_3} & \sigma_{e_3}^2 \end{bmatrix}$$ (3 fixed effects + 6 var/cov = 9 parameters)

- In SAS PROC MIXED, this is done for us automatically:

Null Model Likelihood Ratio Test		
DF	Chi-Square	Pr > ChiSq
5	78.45	<.0001

- Therefore, we find that the UN model fits better

Why the Right Covariance Matrix Matters: Standard Errors and Inferences Made from Fixed Effects

- Fixed effects from the UN model:

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	134.11	4.9820	18	26.92	<.0001
dINTENSITY_W	-46.6111	2.1230	18	-21.96	<.0001
dINTENSITY_R	55.4444	2.2976	18	24.13	<.0001

- Fixed effects from the VC model:

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	134.11	5.1611	54	25.99	<.0001
dINTENSITY_W	-46.6111	7.2988	54	-6.39	<.0001
dINTENSITY_R	55.4444	7.2988	54	7.60	<.0001

What Happens with Differing Models for the Covariances

- The different models for the covariances generally don't change the model for the means (the fixed effects) much
  - Exceptions: unbalanced data

- The standard errors for the fixed effects are derived from the **R** matrix that was estimated:

$$V(\boldsymbol{\beta}) = (\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X})^{-1}$$

  - Putting the wrong $\mathbf{R}$ matrix in the model will lead to the wrong SEs
  - **The wrong SEs will end up giving you inaccurate p-values**
  - Inaccurate p-values will lead to the wrong inferences

- Therefore, the main part of a multivariate model is to determine the appropriate model for the variances
  - This is why we have Repeated Measures (one type of R matrix) and MANOVA (an unstructured R matrix)
  - Unless sample size isn't an issue, the model selected should the model that fits best with the least number of parameters

- Maximum likelihood has made many types of covariance matrices possible

MULTIPLE MODELS FOR COVARIANCE MATRICES

A Multivariate Modeling Demonstration

- To demonstrate the process of finding the best fitting/most parsimonious covariance matrix, we will estimate five models
  1. Variance Components
  2. Variance Components with Heterogeneous Variances
  3. Compound Symmetry
  4. Compound Symmetry with Heterogeneous Variances
  5. Unstructured

- The unstructured model from the previous slides will be the best one can do – but the question remains as to whether any simpler forms would be approximately correct but have fewer parameters

- The choice of a covariance matrix is typically aided by the types of outcomes:
  - Time sensitive? Auto regressive/Toeplitz
  - Same outcome after multiple trials? Unstructured/Compound symmetry
  - Are outcomes region or geography specific? Spatial models

Type 1: Variance Components

- **R** matrix form:

$$\mathbf{R} = \sigma^2\mathbf{I} = \begin{bmatrix} \sigma_e^2 & 0 & 0 \\ 0 & \sigma_e^2 & 0 \\ 0 & 0 & \sigma_e^2 \end{bmatrix}$$

- Estimated **R** matrix:

$$\mathbf{R} = \begin{bmatrix} 479.46 & 0 & 0 \\ 0 & 479.46 & 0 \\ 0 & 0 & 479.46 \end{bmatrix}$$

- Model Fit Statistics:
  - -2 Log L = 486.6
  - Parameters = 4 (3 fixed effects + 1 variances)

Type 2: Heterogeneous Variances/Zero Covariances
TYPE = UN(1) in PROC MIXED

- **R** matrix form:

$$\mathbf{R} = \begin{bmatrix} \sigma^2_{e_1} & 0 & 0 \\ 0 & \sigma^2_{e_2} & 0 \\ 0 & 0 & \sigma^2_{e_3} \end{bmatrix}$$

- Estimated **R** matrix:

$$\mathbf{R} = \begin{bmatrix} 264.36 & 0 & 0 \\ 0 & 446.77 & 0 \\ 0 & 0 & 727.25 \end{bmatrix}$$

- Model Fit Statistics:
  - -2 Log L = 482.1
  - Parameters = 6 (3 fixed effects + 3 variances)

- Model comparison:
  - LRT compared with TYPE=VC: $-2LL = 4.49, df = 2, p = .106$
  - VC is preferred to this model

Type 3: Compound Symmetry [TYPE = CS in PROC MIXED]

- **R** matrix form:

$$\mathbf{R} = \begin{bmatrix} \sigma_e^2 + \sigma & \sigma & \sigma \\ \sigma & \sigma_e^2 + \sigma & \sigma \\ \sigma & \sigma & \sigma_e^2 + \sigma \end{bmatrix}$$

Btw, this is univariate repeated measures ANOVA.

- Estimated **R** matrix:

$$\mathbf{R} = \begin{bmatrix} 479.46 & 409.41 & 409.41 \\ 409.41 & 479.46 & 409.41 \\ 409.41 & 409.41 & 479.46 \end{bmatrix}$$

- Model Fit Statistics:
  - -2 Log L = 435.3
  - Parameters = 5 (3 fixed effects + 1 variances + 1 covariance)

- Model comparison:
  - LRT compared with TYPE=VC: $-2LL = 51.31, df = 1, p < .0001$
  - CS is preferred to VC (so we now use CS as null model)

Type 4: Compound Symmetry/Heterogeneous Variances [TYPE = CSH in PROC MIXED]

- **R** matrix form:

$$\mathbf{R} = \begin{bmatrix} \sigma_{e_1}^2 & \sigma_{e_1}\sigma_{e_2}\rho & \sigma_{e_1}\sigma_{e_3}\rho \\ \sigma_{e_1}\sigma_{e_2}\rho & \sigma_{e_2}^2 & \sigma_{e_2}\sigma_{e_3}\rho \\ \sigma_{e_1}\sigma_{e_3}\rho & \sigma_{e_2}\sigma_{e_3}\rho & \sigma_{e_3}^2 \end{bmatrix}$$

- Estimated **R** matrix:

$$\mathbf{R} = \begin{bmatrix} 217.16 & 310.26 & 403.39 \\ 310.26 & 433.06 & 509.79 \\ 403.39 & 509.79 & 732.06 \end{bmatrix}$$

- Model Fit Statistics:
  - -2 Log L = 415.8
  - Parameters = 7 (3 fixed effects + 3 variances + 1 covariance)

- Model comparison:
  - LRT compared with TYPE=CS: $-2LL = 19.5, df = 2, p < .0001$
  - CSH is preferred to CS (so we now use CSH as null model)

Type 5: Unstructured Covariance Matrix [TYPE = UN in PROC MIXED]

- **R** matrix form:

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

- Estimated **R** matrix:

$$\mathbf{R} = \begin{bmatrix} 264.36 & 315.00 & 373.72 \\ 315.00 & 446.75 & 536.49 \\ 373.72 & 539.49 & 727.25 \end{bmatrix}$$

Btw, this is multivariate (repeated measures) ANOVA.

- Model Fit Statistics:
  - -2 Log L = 408.1
  - Parameters = 9 (3 fixed effects + 3 variances + 3 covariances)

- Model comparison:
  - LRT compared with TYPE=CSH: $-2LL = 7.7, df = 2, p < .021$
  - UN is preferred to CSH – and UN is the winner!

WRAPPING UP

Wrapping Up

- Today's class was our first step into multivariate modeling where multiple outcomes were modeled using a conditional multivariate normal distribution

- Next class: we uncover how adding predictors works for multivariate models
  - How these models relate to classic MANOVA
  - More with the CLASS, ESTIMATE, LSMEANS, (and now) CONTRAST statement, too