Multilevel Models in Matrix Form

Introduction to Multilevel
Models Workshop

University of Georgia:
Institute for Interdisciplinary Research in
Education and Human Development

Covered this Section

- Prerequisites for MLM in matrices:
  - Multivariate normal distribution
  - Matrix algebra
- MLM in matrices

THE MULTIVARIATE NORMAL DISTRIBUTION

Multivariate Normal Distribution

- The generalization of the univariate normal distribution to multiple variables is called the multivariate normal distribution (MVN)
- Many multivariate techniques rely on this distribution in some manner
  - Multilevel/Mixed models

Univariate Normal Distribution

- The univariate normal distribution function is:

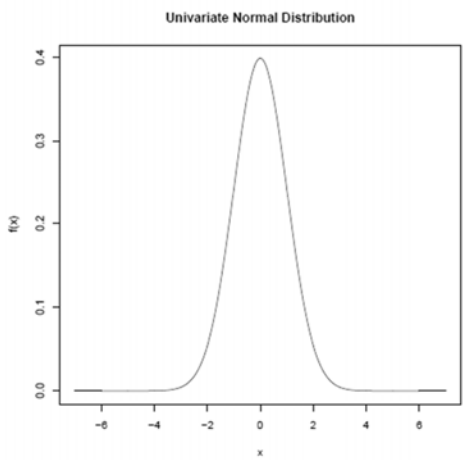
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]$$

- The mean is μ
- The variance is σ^2
- Standard notation for random variables *x* following normal distributions is

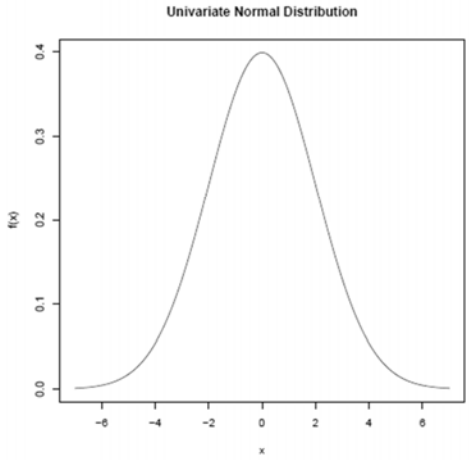
$$x \sim N(\mu, \sigma^2)$$

Univariate Normal Distribution

$N(0,1)$



$N(0,2)$



Multivariate Normal

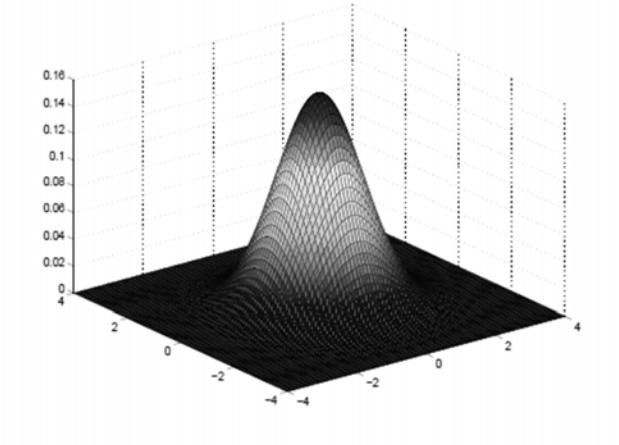
- The multivariate normal distribution function is:

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

- The mean vector is $\boldsymbol{\mu}$
- The covariance matrix is $\boldsymbol{\Sigma}$
- Standard notation for the MVN distribution of p variables is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

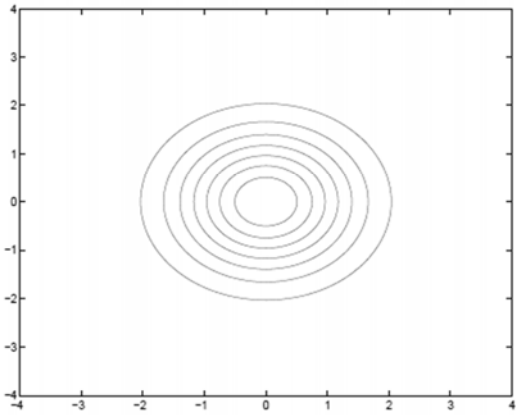
Picturing the Multivariate Normal

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



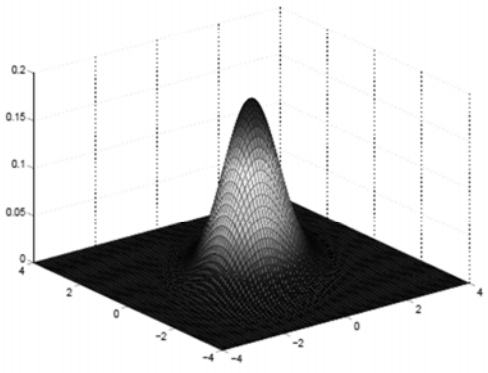
Contour Plot (View From Above)

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

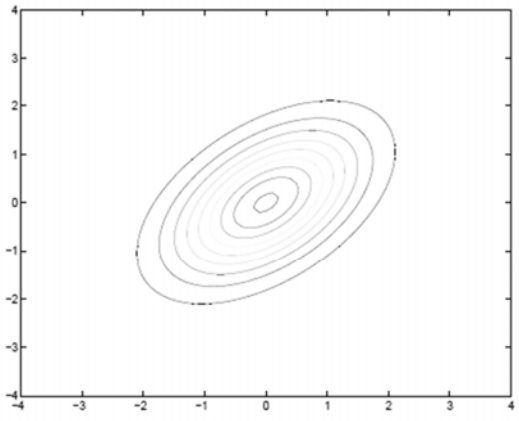


Another Multivariate Normal Plot

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$



MVN Properties

- The MVN distribution has some convenient properties for mixed models
- If **x** has a multivariate normal distribution, then:
  1. Linear combinations of **x** are normally distributed.
  2. All subsets of the components of **x** have an MVN distribution.
  3. Zero covariance implies that the corresponding components are independently distributed.
  4. The conditional distributions of the components are MVN.
     - Especially important for our models

LINEAR MODELS IN MATRICES

Linear Models with Matrices

- Recall our basic linear model (here a regression model) for observation *i* (of *N*):

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + e_i$$

- The equation above can be expressed more compactly by a set of matrices

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

- **Y** is of size (N x 1)
- **X** is of size (N x (1 + k))
- $\boldsymbol{\beta}$ is of size ((1+k) x 1)
- **e** is of size (N x 1)

Unpacking the Equation

$$\underset{\mathbf{Y}}{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix}} = \underset{\mathbf{X}}{\begin{bmatrix} 1 & X_{11} & \cdots & X_{1k} \\ 1 & X_{21} & \cdots & X_{2k} \\ & & \vdots & \\ 1 & X_{N1} & \cdots & X_{Nk} \end{bmatrix}} \underset{\boldsymbol{\beta}}{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}} + \underset{\boldsymbol{e}}{\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}}$$

$$(N\ x\ 1) \qquad (N\ x\ (1+k)) \qquad ((1+k)\ x\ 1) \qquad (N\ x\ 1)$$

For the first observation:

$$Y_1 = \beta_0 + \beta_1 X_{11} + \cdots + \beta_k X_{1k} + e_1$$

Notes on Matrices

- The use of matrices allows for a compact form of the model equation
  - All observations are included

- The matrix of predictors, **X**, has the first column containing all ones
  - Corresponds (multiplies) the intercept β_0
  - Shows how design matrices can be used for in linear models
    - Think about categorical predictors (dummy coding/effect coding)

Linear Model Assumptions

- Recall that we assumed that the error terms were assumed to be
  - Independent
  - Normally distributed $e_i \sim N(0, \sigma_e^2)$

- With matrices, we can now talk about the joint distribution of error terms (for everyone)

$$\mathbf{e} \sim N_N(\mathbf{0}, \sigma_e^2 \mathbf{I}_N)$$

Error Covariance Matrix

- The fixed effects linear model assumes the following structure for the errors:

$$\sigma_e^2 \mathbf{I}_N = \begin{bmatrix} \sigma_e^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_e^2 \end{bmatrix}$$

- In multilevel analyses, this assumption is not valid – so our models introduce terms to relax this assumption

Estimation in Linear Models

- Regression estimates are typically found via least squares (called L_2 estimates)

- In least squares regression, the estimates are found by minimizing the sum of squared errors:

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_k X_{ik})^2$$

- As you could guess, we could do this matrices:

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left(Y_i - \mathbf{x}_i^T \boldsymbol{\beta}\right)^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{e}^T \mathbf{e}$$

The Estimator

- The equation for $\boldsymbol{\beta}$ that minimizes $\mathbf{e}^T\mathbf{e}$ is:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

- The nice thing about this equation is that simultaneously is the MLE (maximizes the likelihood function under normality assumptions)

Model Assumptions

- The conditional distribution of Y has a multivariate normal distribution:
  - Mean vector is the predicted values of Y
  - Covariance matrix is error covariance matrix

$$f(\mathbf{Y}|\mathbf{X}) \sim N_N(\mathbf{X}\boldsymbol{\beta}, \sigma_e^2\mathbf{I}_N)$$

Variance of Estimates

- The covariance matrix of $\boldsymbol{\beta}$ contains useful information regarding the standard errors of the estimates (which are found along the diagonal)

- Under the linear model, this is given by:

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

MULTILEVEL MODELS IN MATRICES (GENERAL LINEAR MIXED MODELS)

Multilevel (Mixed) Models

- The general linear mixed model is given by:

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{Z\gamma} + \mathbf{e}$$

- **Y** is of size (N x 1)
- **X** is of size (N x (1+k))
- $\boldsymbol{\beta}$ is of size ((1+k) x 1)
- **Z** is of size (N x r*g) (r random effects; g groups)
- $\boldsymbol{\gamma}$ is of size (r*g x 1)
- **e** is of size (N x 1)

The New Terms

- The **Z** matrix is analogous to the **X** matrix – it contains the predictors of the *random* effects (i.e., random intercepts, slopes, etc…)
- The $\boldsymbol{\gamma}$ matrix contains the random effects for each observation
- Because of the size of the observations, these matrices are rather large
  - Can be notated differently, though

Z and γ for a Random Intercept

- For a model with a random intercept, this is how **Z** and **γ** appear:

Columns Represent Groups

Columns Represent Type of Effect

Rows Represent Observations

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 & 0 & \dots \\ 1 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \cdots \\ 0 & 1 & 0 & \cdots \end{bmatrix}; \boldsymbol{\gamma} = \begin{bmatrix} \gamma_{10} \\ \gamma_{20} \\ \gamma_{30} \\ \vdots \end{bmatrix}$$

Rows Represent Group Effect Values

Multilevel (Mixed) Model Assumptions

- Assumptions in multilevel (mixed) models involve the random effects and the error terms
- Random effect assumptions:
  - Multivariate Normal (across *r* random effects)
  - Mean Vector **0**; Covariance Matrix **G** (block diagonal within a group)

$$\boldsymbol{\gamma} \sim N_r(\mathbf{0}, \mathbf{G})$$

- Error term assumptions
  - Multivariate Normal (within a group)
  - Mean Vector **0**; Covariance Matrix **R**

$$\mathbf{e} \sim N_N(\mathbf{0}, \mathbf{R})$$

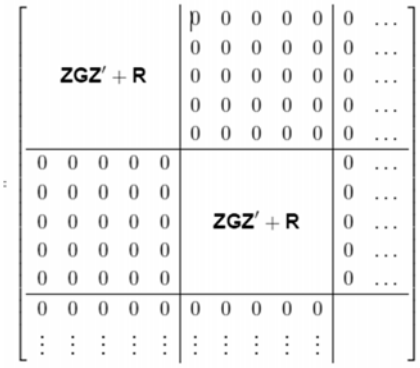
Model Assumptions

- The conditional distribution of Y has a multivariate normal distribution:
  - Mean vector is the predicted values of Y
  - Covariance matrix is combination of random effect and error term covariance matrices
    - Allows for correlated observations

$$f(\mathbf{Y}|\mathbf{X},\mathbf{Z}) \sim N_N(\mathbf{X\beta}, \mathbf{ZGZ}^T + \mathbf{R})$$

New Covariance Matrix

- Because of the grouping structure of data, the new covariance matrix is block-diagonal
- Blocks represent the covariance matrix for a group/cluster of observations



Model Estimation

- Because of the inclusion of random effects (which are not directly observable), the model no longer has a single estimation equation
- Rather, we now must use an iterative process to estimate model parameters
- Two estimators are commonly used: maximum likelihood (ML) and residual maximum likelihood (REML)
  - I will introduce ML first then REML

ML Estimation of Mixed Models

- The goal in ML estimation is to pick a set of parameters that maximize the likelihood function
  - Typically the log-likelihood is used
  - Here, we have to know $\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{G}, \mathbf{R}$
    - $\boldsymbol{\gamma}$ isn't a part of the function below
- The log-likelihood function is the log of the model-assumed MVN:

$$N_N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R})$$

Simplifying Things

- Because of the wonders of math, we can use a technique called *estimated generalized least squares*
  - Use some method to find **G** and **R**: $\widehat{\mathbf{G}}$ and $\widehat{\mathbf{R}}$
  - Given **G** and **R**, we can find $\boldsymbol{\beta}$
  - Here, we will define $\widehat{\mathbf{V}} = \mathbf{Z}\widehat{\mathbf{G}}\mathbf{Z}^T + \widehat{\mathbf{R}}$
- Specifically:

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\widehat{\mathbf{V}}\mathbf{X}\right)^{-1}\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{Y}$$

The ML Log Likelihood

- The goal is to pick **G** and **R** and then substitute them into the log likelihood function, producing a log likelihood value
- Picking **G** and **R** can be done using Newton-Raphson (as is done in SAS)
- The function value is:

$$l(\mathbf{G},\mathbf{R}) = -\frac{1}{2}\log\left|\widehat{\mathbf{V}}\right| - \frac{1}{2}\mathbf{r}^T\widehat{\mathbf{V}}^{-1}\mathbf{r} - \frac{\mathrm{n}}{2}\log(2\pi)$$

Where:

$$\mathbf{r} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{Y} - \mathbf{X}\left(\mathbf{X}^T\widehat{\mathbf{V}}\mathbf{X}\right)^{-1}\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{Y}$$

Issues with ML Estimates

- ML estimation is a common choice and performs well when sample sizes are large
- However, estimates of the variances will be biased
  - Similar to basic statistics phenomena of using N versus N-1 in the variance/standard deviation
- Therefore, the residual ML estimator was developed
  - Called REML

REML Estimator

- The REML estimator maximizes the likelihood of the residuals
- The likelihood function comes from stating the likelihood of the data as a function of the likelihood of the estimated fixed effects and the residuals
- Here, we take the estimated residuals to be

$$\hat{\mathbf{e}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

- Where $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X}\right)^{-1}\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{Y}$

Deriving REML

- Because **Y** is multivariate normal, $\widehat{\boldsymbol{\beta}}$ and $\hat{\mathbf{e}}$ are linear functions of **Y** that are:
  - Normally distributed (see properties of MVN)
  - Independent
- Therefore, with independence we can re-express the likelihood of **Y** as a product of $\widehat{\boldsymbol{\beta}}$ and $\hat{\mathbf{e}}$

$$L\left(\mathbf{Y}\middle|\widehat{\mathbf{V}}\right) = L\left(\widehat{\boldsymbol{\beta}}\middle|\widehat{\mathbf{V}}\right)L\left(\hat{\mathbf{e}}\middle|\widehat{\mathbf{V}}\right)$$

More Deriving REML

- Further, due to the consistency of the estimates, we know that

$$\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \left(\mathbf{X}^T\widehat{\mathbf{V}}^{-\mathbf{1}}\mathbf{X}\right)^{-\mathbf{1}}\right)$$

- Therefore, it is now our goal to maximize the log-likelihood of the residuals, or $L(\hat{\mathbf{e}}|\ \widehat{\mathbf{V}})$

Step 1: Taking the Log

- We now take the log of our original likelihood function:

$$L(\mathbf{Y}|\widehat{\mathbf{V}}) = L(\widehat{\boldsymbol{\beta}}|\widehat{\mathbf{V}})L(\hat{\mathbf{e}}|\widehat{\mathbf{V}})$$

- Yielding:

$$\log\left(L(\mathbf{Y}|\widehat{\mathbf{V}})\right) = \log\left(L(\widehat{\boldsymbol{\beta}}|\widehat{\mathbf{V}})\right) + \log\left(L(\hat{\mathbf{e}}|\widehat{\mathbf{V}})\right)$$

- Which gives us:

$$\log\left(L(\hat{\mathbf{e}}|\widehat{\mathbf{V}})\right) = \log\left(L(\mathbf{Y}|\widehat{\mathbf{V}})\right) - \log\left(L(\widehat{\boldsymbol{\beta}}|\widehat{\mathbf{V}})\right)$$

Step 2:

- We know that $\mathbf{Y} \sim N_N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R})$ and $\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \left(\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\right)$

- We can then put the MVN associated with each into our log likelihood of the residual

$$\log\left(L(\hat{\mathbf{e}}|\widehat{\mathbf{V}})\right) = \log\left(L(\mathbf{Y}|\widehat{\mathbf{V}})\right) - \log\left(L(\widehat{\boldsymbol{\beta}}|\widehat{\mathbf{V}})\right)$$

Even More…

$$\log\left(L\left(\hat{\mathbf{e}}\middle|\widehat{\mathbf{V}}\right)\right) = \log\left(L\left(\mathbf{Y}\middle|\widehat{\mathbf{V}}\right)\right) - \log\left(L\left(\widehat{\boldsymbol{\beta}}\middle|\widehat{\mathbf{V}}\right)\right)$$

$$= -\frac{1}{2}\Big[\log\left|\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{X}\right| + \log\left|\widehat{\mathbf{V}}\right| + (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T\widehat{\mathbf{V}}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) - \left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^T\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\Big]$$

Here:

$$(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T\widehat{\mathbf{V}}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) = \left(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)^T\widehat{\mathbf{V}}^{-1}\left(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right) + \left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^T\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)$$

Meaning we can cancel the last term….

The REML Log Likelihood

- After all the slides before, we can now present the REML log likelihood:

$$\log\left(L\left(\hat{\mathbf{e}}\middle|\widehat{\mathbf{V}}\right)\right) = -\frac{1}{2}\left[\log\left|\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{X}\right| + \log\left|\widehat{\mathbf{V}}\right| + \hat{\mathbf{e}}^T\hat{\mathbf{e}} - \frac{n-p}{2}\log(2\pi)\right]$$

Uses of ML and REML

- ML can be used for deviance tests when the fixed effects are the same or are different
- REML can be used for deviance tests when the fixed effects are the same only
  - Residuals change when the fixed effects change

Final Thoughts

- We discussed the matrix form of linear models with mixed effects
  - Multilevel models
- The matrix form can be useful for reading about these models in papers and presentations
- This section was meant to be an introduction to the technical side of the modeling framework
  - Much more time can be spent on just this alone