Lecture 2:
Linear and Mixed Models

Bruce Walsh lecture notes

Introduction to Mixed Models

SISG (Module 12), Seattle

22 – 24 July 2020

First PAUSE at slide 9

The simple linear regression

Consider a simple linear regression,

$$y_i = a + b\,x_i + e_i \quad 1 \le i \le n$$

We can express in matrix form as follows. First, define

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} a \\ b \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

Then from the definitions of matrix multiplication and addition,

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

Predicted values and residuals

$$y_i = a + b\,x_i + e_i$$

$$y_i = \widehat{y_i} + e_i$$

The predicted value from the model is

$$\widehat{y_i} = a + b\,x_i$$

The model residual is the the deviation from observed and predicted values

$$\widehat{e_i} = y_i - \widehat{y_i}$$

In matrix form

$$\widehat{\mathbf{y}} = \mathbf{X}\,\widehat{\mathbf{b}}$$

$$\widehat{\mathbf{e}} = \mathbf{y} - \widehat{\mathbf{y}}$$

Simple 2-factor ANOVA

- Suppose you have three observations (4,5,7) from males and four observations from females (3,5,6,8)
  - As we will see, *matrices easily handle unbalanced designs*!
- The simple ANOVA for this model is
- $y_{ij} = s_i + e_{ij}$,
  - for males (i=1) and females (i=2)

In matrix form

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

where

$$\mathbf{y} = \begin{pmatrix} 4 \\ 5 \\ 7 \\ 3 \\ 5 \\ 6 \\ 8 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} s_1 \\ s_2 \end{pmatrix}$$

Quick Overview of the Major Points

The general linear model can be written as

$$y = X\beta + e$$

- y = vector of observed dependent values

- X = Design matrix: observations of the variables in the assumed linear model

- β = vector of unknown parameters to estimate

- e = vector of residuals (deviation from model fit),
  $e = y - X\beta$

$y = X\beta + e$

Solution to β depends on the *__covariance structure__*
(= covariance matrix) of the vector **e** of residuals

Ordinary least squares (OLS)

- OLS: $e \sim MVN(\mathbf{0}, \sigma^2 I)$
- Residuals are homoscedastic and uncorrelated,
  so that we can write the cov matrix of **e** as $Cov(e) = \sigma^2 I$
- the OLS estimate, $OLS(\beta) = (X^T X)^{-1} X^T y$

Generalized least squares (GLS)

- GLS: $e \sim MVN(\mathbf{0}, V)$
- Residuals are heteroscedastic and/or dependent,
- $GLS(\beta) = (X^T V^{-1} X)^{-1} X^T V^{-1} y$

BLUE

- Both the OLS and GLS solutions are also called the Best Linear Unbiased Estimator (or BLUE for short)
- Whether the OLS or GLS form is used depends on the assumed covariance structure for the residuals
  - Special case of Var($\mathbf{e}$) = $\sigma_e^2 \mathbf{I}$ -- OLS
  - All others, i.e., Var($\mathbf{e}$) = $\mathbf{R}$ -- GLS

PAUSE

- Most common statistical problems can be expressed as special cases of the General Linear Model (GLM)

- These follow from the definition of matrix multiplication.

- The predicted value for observation i is just the inner product of the ith row of **X** on the column vector β of model parameters

- Assumptions on the covariance structure of the residuals (OLS or GLS) are required for a solution

- Next PAUSE is at slide 26

Linear Models

One tries to explain a dependent variable y as a linear function of a number of independent (or predictor) variables.

A multiple regression is a typical linear model,

$$y = \mu + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_x + e$$

Here e is the residual, or deviation between the true value observed and the value predicted by the linear model.

The (partial) regression coefficients are interpreted as follows: a unit change in x_i while holding all other variables constant results in a change of β_i in y

Linear Models

As with a univariate regression ($y = a + bx + e$), the model parameters are typically chosen by least squares, wherein they are chosen to minimize the sum of squared residuals, $\Sigma\, e_i^2 = e^T e$

This unweighted sum of squared residuals assumes an OLS error structure, so all residuals are equally weighted (homoscedastic) and uncorrelated

If the residuals differ in variances and/or some are correlated (GLS conditions), then we need to minimize the **weighted sum** $e^T V^{-1} e$, which removes correlations and gives all residuals equal variance.

Linear Models in Matrix Form

Suppose we have 3 variables in a multiple regression, with four (y,x) vectors of observations.

$$y_i = \mu + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$$

In matrix form, $\quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ 1 & x_{31} & x_{32} & x_{33} \\ 1 & x_{41} & x_{42} & x_{43} \end{pmatrix} \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

The design matrix X. Details of both the experimental design and the observed values of the predictor variables all reside solely in X

Rank of the design matrix

- With n observations and p unknowns, X is an n x p matrix, so that X^TX is p x p
- Thus, at most X can provide unique estimates for up to p < n parameters
- The rank of X is the number of independent rows of X. If X is of full rank, then rank = p
- A parameter is said to be estimable if we can provide a unique estimate of it. If the rank of X is k < p, then exactly k parameters are estimable (some as linear combinations, e.g. $\beta_1 - 3\beta_3 = 4$)
- if $\det(X^TX) = 0$, then X is not of full rank
- Number of nonzero eigenvalues of X^TX gives the rank of X.

Experimental design and X

- The structure of X determines not only which parameters are estimable, but also the expected sample variances, as $Var(\beta) = k(X^TX)^{-1}$

- Experimental design determines the structure of X before an experiment (of course, missing data almost always means the final X is different form the proposed X)

- Different criteria used for an optimal design. Let $V = (X^TX)^{-1}$. The idea is to chose a design for X given the constraints of the experiment that:
  - A-optimality: minimizes $tr(V)$
  - D-optimality: minimizes $det(V)$
  - E-optimality: minimizes leading eigenvalue of V

Ordinary Least Squares (OLS)

When the covariance structure of the residuals has a certain form, we solve for the vector β using OLS

If residuals follow a MVN distribution, OLS = ML solution

If the residuals are homoscedastic and uncorrelated, $\sigma^2(e_i) = \sigma_e^2$, $\sigma(e_i, e_j) = 0$. Hence, each residual is equally weighted,

Sum of squared residuals can be written as

$$\sum_{i=1}^{n} \widehat{e}_i^2 = \widehat{\mathbf{e}}^T \widehat{\mathbf{e}} = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

Predicted value of the y's

Ordinary Least Squares (OLS)

$$\sum_{i=1}^{n} \widehat{e}_i^2 = \widehat{\mathbf{e}}^T \widehat{\mathbf{e}} = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

Taking (matrix) derivatives shows this is minimized by

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

This is the OLS estimate of the vector β

The variance-covariance estimate for the sample estimates is

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

The ij-th element gives the covariance between the estimates of β_i and β_j.

16

Sample Variances/Covariances

The residual variance can be estimated as

$$\widehat{\sigma_e^2} = \frac{1}{n - \text{rank}(X)} \sum_{i=1}^{n} \widehat{e}_i^2$$

The estimated residual variance can be substituted into

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

To give an approximation for the sampling variance and covariances of our estimates.

Confidence intervals follow since the vector of estimates ~ MVN(β, $\mathbf{V}_\beta$)

Example: Regression Through the Origin

$$y_i = \beta x_i + e_i$$

Here
$$\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \boldsymbol{\beta} = (\beta)$$

$$\mathbf{X}^T\mathbf{X} = \sum_{i=1}^{n} x_i^2 \qquad \mathbf{X}^T\mathbf{y} = \sum_{i=1}^{n} x_i\, y_i$$

$$\boxed{\begin{array}{l} \beta = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \dfrac{\sum x_i\, y_i}{\sum x_i^2} \\[2em] \sigma^2(\beta) = \dfrac{1}{n-1}\dfrac{\sum(y_i - \beta x_i)^2}{\sum x_i^2} \end{array}}$$

$$\sigma^2(b) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\sigma_e^2 = \dfrac{\sigma_e^2}{\sum x_i^2}$$

$$\sigma_e^2 = \dfrac{1}{n-1}\sum(y_i - \beta x_i)^2$$

Polynomial Regressions

GLM can easily handle any function of the observed predictor variables, provided the parameters to estimate are still linear, e.g. $y = \alpha + \beta_1 f(x) + \beta_2 g(x) + \cdots + e$

Quadratic regression:

$$y_i = \alpha + \beta_1\, x_i + \beta_2\, x_i^2 + e_i$$

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}$$

Interaction Effects

Interaction terms (e.g. sex x age) are handled similarly

$$y_i = \alpha + \beta_1\, x_{i1} + \beta_2\, x_{i2} + \beta_3\, x_{i1}x_{i2} + e_i$$

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{11}x_{12} \\ 1 & x_{21} & x_{22} & x_{21}x_{22} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n1}x_{n2} \end{pmatrix}$$

With x_1 held constant, a unit change in x_2 changes y by $\beta_2 + \beta_3 x_1$ (i.e., the slope in x_2 depends on the current value of x_1)

Likewise, a unit change in x_1 changes y by $\beta_1 + \beta_3 x_2$

The GLM lets you build your own model!

- Suppose you want a quadratic regression forced through the origin where the slope of the quadratic term can vary over the sexes (pollen vs. seed parents)

- $y_i = \beta_1 x_i + \beta_2 x_i^2 + \beta_3 s_i x_i^2$

- s_i is an indicator (0/1) variable for the sex (0 = male, 1 = female).
  - Male slope = β_2,
  - Female slope = $\beta_2 + \beta_3$

Generalized Least Squares (GLS)

Suppose the residuals no longer have the same variance (i.e., display heteroscedasticity). Clearly we do not wish to minimize the *unweighted* sum of squared residuals, because those residuals with smaller variance should receive more weight.

Likewise in the event the residuals are correlated, we also wish to take this into account (i.e., perform a suitable transformation to remove the correlations) before minimizing the sum of squares.

Either of the above settings leads to a GLS solution in place of an OLS solution.

In the GLS setting, the covariance matrix for the vector e of residuals is written as R where
$R_{ij} = \sigma(e_i, e_j)$

The linear model becomes $y = X\beta + e$, $cov(e) = R$

The GLS solution for β is

$$b = \left(X^T R^{-1} X\right)^{-1} X^T R^{-1} y$$

The variance-covariance of the estimated model parameters is given by

$$V_b = \left(X^T R^{-1} X\right)^{-1} \sigma_e^2$$

Model diagnostics

- It's all about the residuals
- Plot the residuals
  - Quick and easy screen for outliers
  - Plot y or yhat vs e
- Test for normality among estimated residuals
  - Q-Q plot
  - Wilk-Shapiro test
  - If non-normal, try transformations, such as log

OLS, GLS summary

	OLS	GLS
Assumed distribution of residuals	$\mathbf{e} \sim (\mathbf{0}, \sigma_e^2 \mathbf{I})$	$\mathbf{e} \sim (\mathbf{0}, \mathbf{V})$
Least-squares estimator of β	$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$	$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$
$\text{Var}(\widehat{\boldsymbol{\beta}})$	$(\mathbf{X}^T\mathbf{X})^{-1}\sigma_e^2$	$(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$
Predicted values, $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$	$\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$	$\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$
$\text{Var}(\widehat{\mathbf{y}})$	$\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma_e^2$	$\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T$

PAUSE

– Reviewed features of OLS vs GLS

– Presented examples of constructing GLMs

– Estimability, rank of X, eigenvalues of $X^T X$

– Model diagnostics are all about residuals

– Next pause at slide 35

Fixed vs. Random Effects

In linear models are are trying to accomplish two goals: estimation the values of model parameters and estimate any appropriate variances.

For example, in the simplest regression model, $y = \alpha + \beta x + e$, we estimate the values for α and β and also the variance of e. We, of course, can also estimate the $e_i = y_i - (\alpha + \beta x_i)$

Note that α/β are *fixed constants* are we trying to estimate (fixed factors or fixed effects), while the e_i values are drawn from some probability distribution (typically Normal with mean 0, variance σ^2_e). The e_i are random effects.

This distinction between fixed and random effects is extremely important in terms of how we analyzed a model. If a parameter is a fixed constant we wish to estimate, it is a fixed effect. If a parameter is drawn from some probability distribution and we are trying to make inferences on either the distribution and/or specific realizations from this distribution, it is a random effect.

We generally speak of estimating fixed factors (BLUE) and predicting random effects (BLUP -- best linear unbiased Predictor)

"Mixed" models (MM) contain both fixed and random factors

$$y = Xb + Zu + e, \quad u \sim MVN(0,R), \; e \sim MVN(0,\sigma^2_e I)$$

Key: need to specify covariance structures for MM

Random effects models

- It is often useful to treat certain effects as random, as opposed to fixed
    - Suppose we have k effects. If we treat these as fixed, we lose k degrees of freedom
    - If we assume each of the k realizations are drawn from a normal with mean zero and unknown variance, only one degree of freedom lost --- that for estimating the variance
        - We can then predict the values of the k realizations

Environmental effects

- Consider yield data measured over several years in a series of plots.
- Standard to treat year-to-year variation at a specific site as being random effects
- Often the plot effects (mean value over years) are also treated as random.
- For example, consider plants group in growing region i, location j within that region, and year (season) k for that location-region effect
  - $E = R_i + L_{ij} + e_{ijk}$
  - Typically R can be a fixed effect, while L and e are random effects, $L_{ik} \sim N(0, \sigma^2_L)$ and $e_{ikj} \sim N(0, \sigma^2_e)$

Random models

- With a random model, one is assuming that all "levels" of a factor are not observed. Rather, some subset of values are drawn from some underlying distribution
  - For example, year to year variation in rainfall at a location. Each year is a random sample from the long-term distribution of rainfall values
  - Typically, assume a functional form for this underlying distribution (e.g., normal with mean 0) and then use observations to estimate the distribution parameters (here, the variance)

Random models (cont)

- Key feature:
  - Only one degree of freedom used (estimate of the variance)
  - Using the fixed effects and the estimated underlying distribution parameters, one then <u>predicts</u> the actual realizations of the individual values (i.e., the year effects)
  - Assumption: the covariance structure among the individual realizations of the realized effects. If only a variance is assume, this implies they are independent. If they are assumed to be correlated, this structure must be estimated.

Random models

- Let's go back to treating yearly effects as random
- If assume these are uncorrelated, only use one degree of freedom, but makes assumptions about covariance structure
  - Standard: Uncorrelated
  - Option: some sort of autocorrelation process, say with a yearly decay of r (must also be estimated)
- Conversely, could all be treated as fixed, but would use k degrees of freedom for k years, but no assumptions on their relationships (covariance structure)

Identifiability

- Recall that a fixed effect is said to be estimable if we can obtain a unique estimate for it (either because X is of full rank or when using a generalized inverse it returns a unique estimate)
  - Lack of estimable arises because the experiment design confounds effects
- The analogous term for random models is identifiability
  - The variance components have unique estimates

PAUSE

– Distinction between fixed and random effects

 • Often use random effects for convenience to avoid losing degrees of freedom for parameters that you really don't care about, but still need to include in the model.

 • This requires assumptions (i.e., typically assumes independent draws)

– Next PAUSE at slide 40

The general mixed model

Vector of fixed effects (to be estimated), e.g., year, sex and age effects

Vector of observations (phenotypes)

Incidence matrix for random effects

$$y = X\beta + Zu + e$$

Vector of residual errors (random effects)

Incidence matrix for fixed effects

Vector of random effects, such as individual Breeding values (to be estimated)

The general mixed model

Vector of fixed effects

Vector of
observations
(phenotypes)

Incidence matrix for random effects

$$y = X\beta + Zu + e$$

Vector of residual errors

Incidence
matrix for
fixed effects

Vector of random
effects

Observe y, X, Z.

Estimate fixed effects β

Estimate random effects u, e

Means & Variances for $y = X\beta + Zu + e$

Means: $E(u) = E(e) = 0$, $E(y) = X\beta$

Variances:

Let R be the covariance matrix for the residuals. We typically assume $R = \sigma^2_e * I$

Let G be the covariance matrix for the vector u of random effects

The covariance matrix for y becomes
$V = ZGZ^T + R$

Hence, $y \sim MVN(X\beta, V)$

Mean $X\beta$ due to fixed effects
Variance V due to random effects

Estimating fixed Effects & Predicting Random Effects

For a mixed model, we observe y, X, and Z

β, u, R, and G are generally unknown

Two complementary estimation issues

(i) Estimation of **β** and **u**

$$\hat{\beta} = \left(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y}$$ Estimation of fixed effects

BLUE = Best Linear Unbiased Estimator

$$\hat{\mathbf{u}} = \mathbf{G} \mathbf{Z}^T \mathbf{V}^{-1} \left(\mathbf{y} - \mathbf{X}\hat{\beta} \right)$$ Prediction of random effects

BLUP = Best Linear Unbiased Predictor

Recall $V = ZGZ^T + R$

PAUSE

– Mixed models are the logical extensions of GLMs by adding additional vectors of random effects beyond the residual

– As with residuals, their solution requires assumptions about the covariance structure of the random effects

– Last slide is 43

Different statistical models

- GLM = general linear model
  - OLS ordinary least squares: e ~ MVN(0,cI)
  - GLS generalized least squares: e ~ MVN(0,R)
- Mixed models
  - Both fixed and random effects (beyond the residual)
- Mixture models
  - A weighted mixture of distributions
- General<u>ized</u> linear models
  - Nonlinear functions, non-normality

Mixture models

- Under a mixture model, an observation potentially comes from one of several different distributions, so that the density function is $\pi_1\phi_1 + \pi_2\phi_2 + \pi_3\phi_3$
  - The mixture proportions π_i sum to one
  - The ϕ_i represent different distribution, e.g., normal with mean μ_i and variance σ^2

- Mixture models come up in QTL mapping -- an individual could have QTL genotype QQ, Qq, or qq
  - See Lynch & Walsh Chapter 13

- They also come up in codon models of evolution, were a site may be neutral, deleterious, or advantageous, each with a different distribution of selection coefficients
  - See Walsh & Lynch (volume 2A website), Chapters 10,11

Generalized linear models

The **Generalized Linear Model** (note the **ized** ending) takes this a step further by assuming for some monotonic function g, that

$$E[\, y_i \,] = g \left(\mu + \sum_{k=1}^{n} \beta_k x_{ik} \right) \qquad (2)$$

In particular, taking the inverse g^{-1} of the function g returns a linear model, with

$$g^{-1}\left(E[\, y_i \,]\right) = \mu + \sum_{k=1}^{n} \beta_k x_{ik} \qquad (3)$$

The function f with the property that expresses the expected value of the response variable as a linear function of the predictor variables, i.e.,

$$f\left(E[\, y_i \,]\right) = \mu + \sum_{k=1}^{n} \beta_k x_{ik}$$

is called the **link function** of the particular generalized linear model.

Typically assume non-normal distribution for residuals, e.g., Poisson, binomial, gamma, etc