Lecture 4:
Linear and Mixed Models

Bruce Walsh lecture notes
Tucson Winter Institute
9 - 11 Jan 2013

Quick Review of the Major Points

The general linear model can be written as

$$y = X\beta + e$$

- $\mathbf{y}$ = vector of observed dependent values
- $\mathbf{X}$ = Design matrix: observations of the variables in the assumed linear model
- β = vector of unknown parameters to estimate
- $\mathbf{e}$ = vector of residuals (deviation from model fit), $\mathbf{e} = \mathbf{y}-\mathbf{X}\ \beta$

$y = X\beta + e$

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

Ordinary least squares (OLS)

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

Generalized least squares (GLS)

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

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

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, $\Sigma\, e_i^2$

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 $\mathbf{e}^T\mathbf{V}^{-1}\mathbf{e}$, which removes correlations and gives all residuals equal variance.

Predictor and Indicator Variables

Suppose we measuring the offspring of p sires. One linear model would be

$$y_{ij} = \mu + s_i + e_{ij}$$

y_{ij} = trait value of offspring j from sire i

μ = overall mean. This term is included to give the s_i terms a mean value of zero, i.e., they are expressed as deviations from the mean

s_i = The effect for sire i (the mean of its offspring). Recall that variance in the s_i estimates Cov(half sibs) = $V_A/4$

e_{ij} = The deviation of the jth offspring from the family mean of sire i. The variance of the e's estimates the within-family variance.

Predictor and Indicator Variables

In a regression, the predictor variables are typically continuous, although they need not be.

$$y_{ij} = \mu + s_i + e_{ij}$$

Note that the predictor variables here are the s_i, (the value associated with sire i) something that we are trying to estimate

We can write this in linear model form, $y_{ij} = \mu + \Sigma_k \, x_{ik} s_i + e_{ij}$, by using indicator variables

$$x_{ik} = \begin{cases} 1 & \text{if sire } k = i \\ 0 & \text{otherwise} \end{cases}$$

Models consisting entirely of indicator variables are typically called **ANOVA**, or **analysis of variance models**

Models that contain no indicator variables (other than for the mean), but rather consist of observed value of continuous or discrete values are typically called **regression models**

Both are special cases of the **General Linear Model** (or **GLM**)

$$y_{ijk} = \mu + s_i + d_{ij} + \beta x_{ijk} + e_{ijk}$$

Example: Nested half sib/full sib design with an age correction β on the trait

Example: Nested half sib/full sib design with an age correction β on the trait

ANOVA model

$$y_{ijk} = \mu + s_i + d_{ij} + \beta x_{ijk} + e_{ijk}$$

Regression model

s_i = effect of sire i

d_{ij} = effect of dam j crossed to sire i

x_{ijk} = age of the kth offspring from i x j cross

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, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$

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

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

In-class Exercise

Suppose you measure height and sprint speed for five individuals, with heights (x) of 9, 10, 11, 12, 13 and associated sprint speeds (y) of 60, 138, 131, 170, 221

1) Write in matrix form (i.e, the design matrix X and vector β of unknowns) the following models

- $y = bx$
- $y = a + bx$
- $y = bx^2$
- $y = a + bx + cx^2$

2) Using the X and y associated with these models, compute the OLS BLUE, $\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ for each

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(β) = 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.

Sample Variances/Covariances

The residual variance can be estimated as

$$\widehat{\sigma_e^2} = \frac{1}{n - \mathrm{rank}(X)} \sum_{i=1}^{n} \widehat{e_i^2}$$

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

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

$$\boxed{\begin{aligned} \beta &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \frac{\sum x_i\, y_i}{\sum x_i^2} \\ \sigma^2(\beta) &= \frac{1}{n-1}\,\frac{\sum (y_i - \beta x_i)^2}{\sum x_i^2} \end{aligned}}$$

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

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

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

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

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

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 $y = X\beta + e$, $\mathrm{cov}(e) = R$

The GLS solution for β is

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

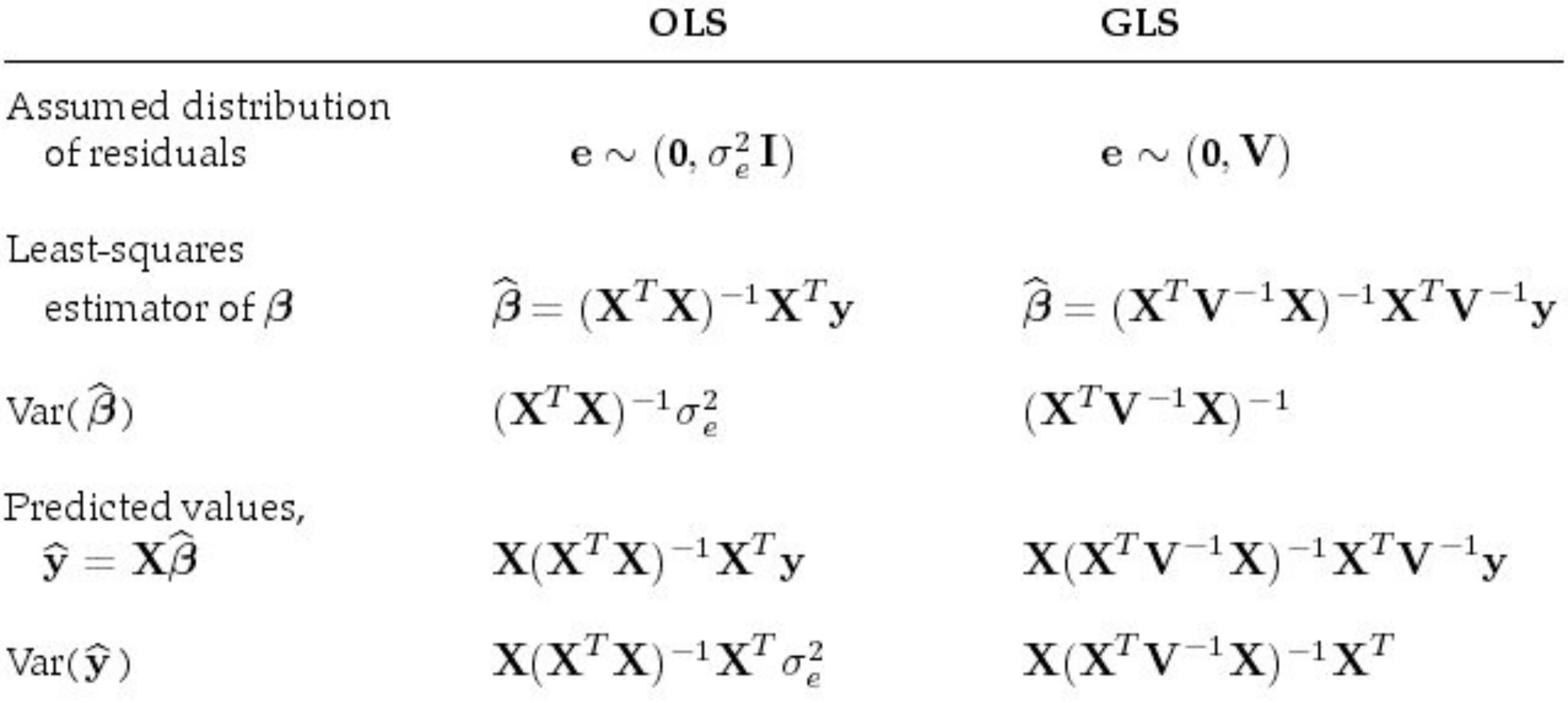
The variance-covariance of the estimated model parameters is given by

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

Model diagnostics

- **It's all about the residuals**
- Plot the residuals
  - Quick and easy screen for outliers
- 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 $\boldsymbol{\beta}$	$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$	$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$
Var($\widehat{\boldsymbol{\beta}}$)	$(\mathbf{X}^T\mathbf{X})^{-1}\sigma_e^2$	$(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$
Predicted values, $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$	$\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$	$\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$
Var($\widehat{\mathbf{y}}$)	$\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma_e^2$	$\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T$

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

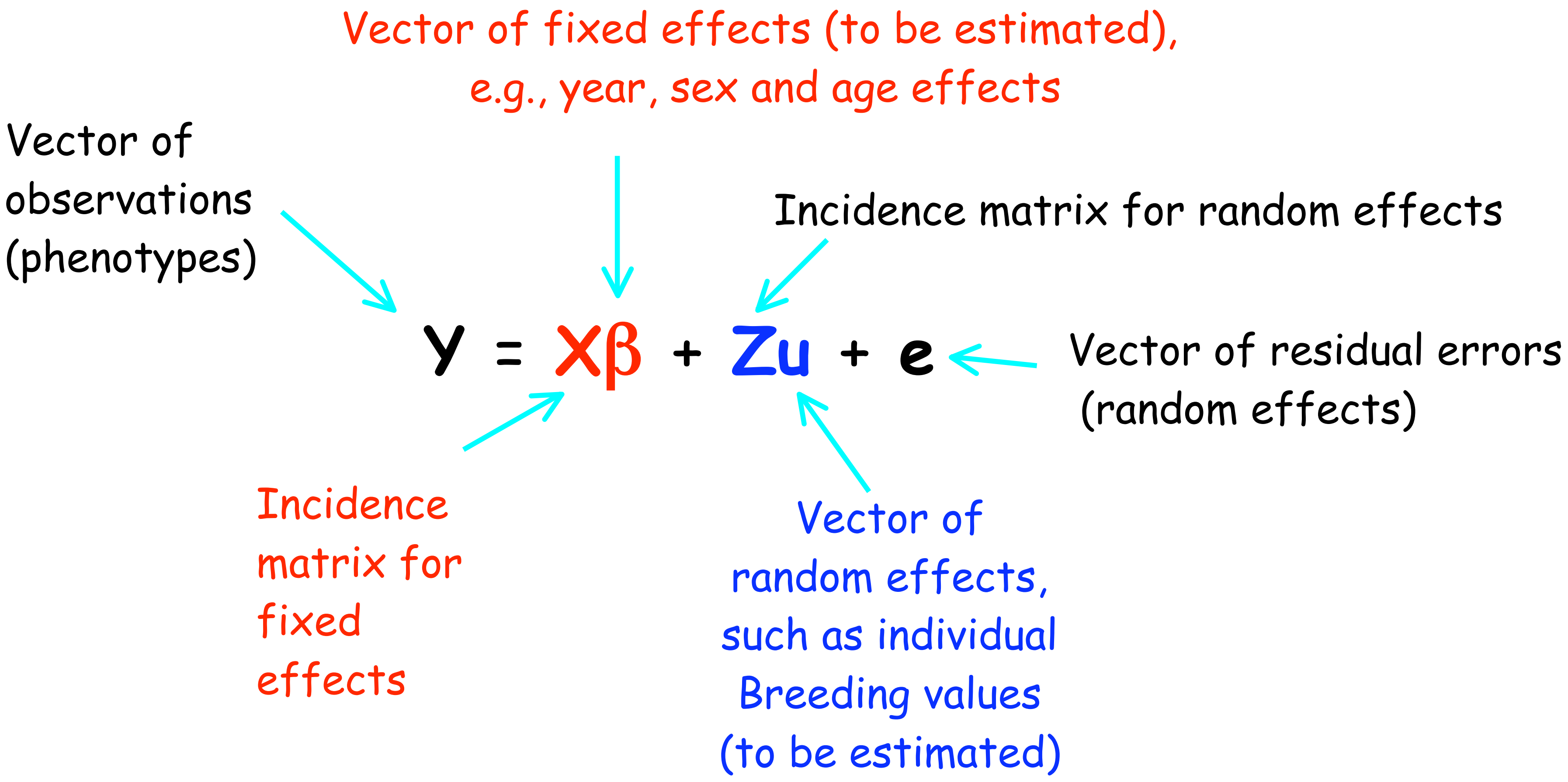
Environmental effects

- 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 k within that region, and year (season) k for that location-region effect
  - $E = R_i + L_{ik} + e_{ijk}$
  - Typically R can be a fixed effect, while L and e are random effects, $L_{ik} \sim N(0,\sigma^2_L)$ and $e_{ikj} \sim N(0,\sigma^2_e)$

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

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

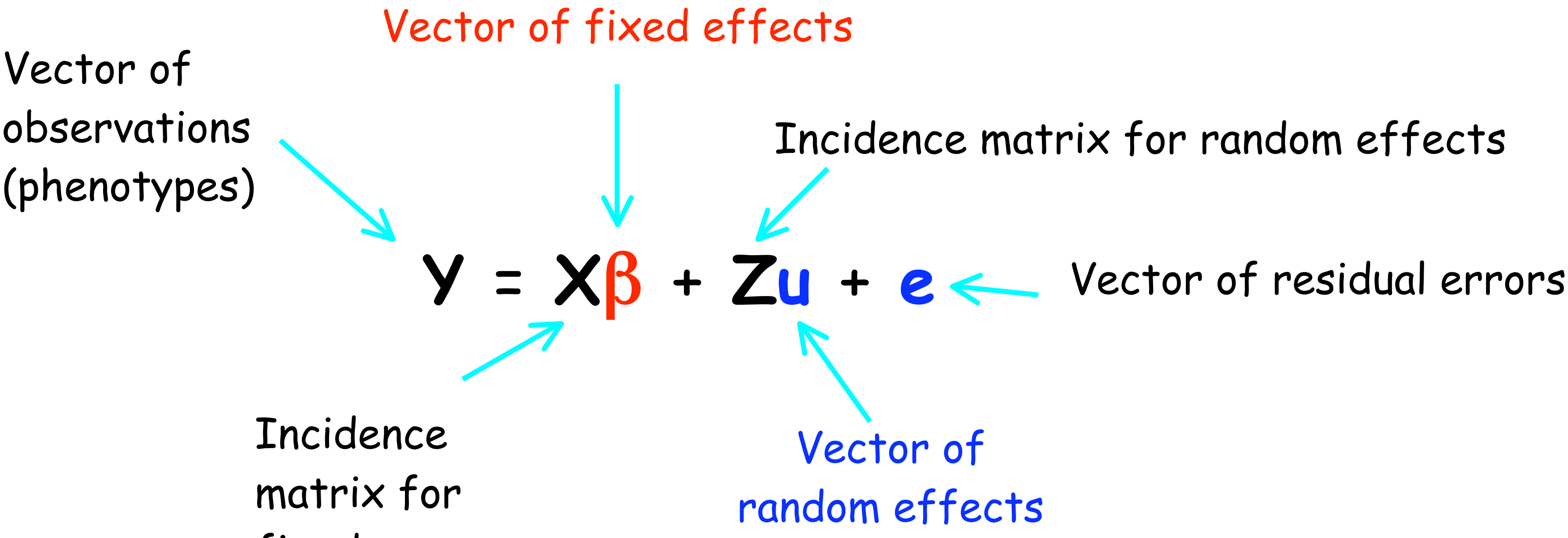
$$y = X\beta + Zu + e$$

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

Vector of residual errors

Incidence matrix for fixed effects

Vector of random effects

Observe $\mathbf{y}$, $\mathbf{X}$, $\mathbf{Z}$.

Estimate fixed effects β

Estimate random effects $\mathbf{u}$, $\mathbf{e}$

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

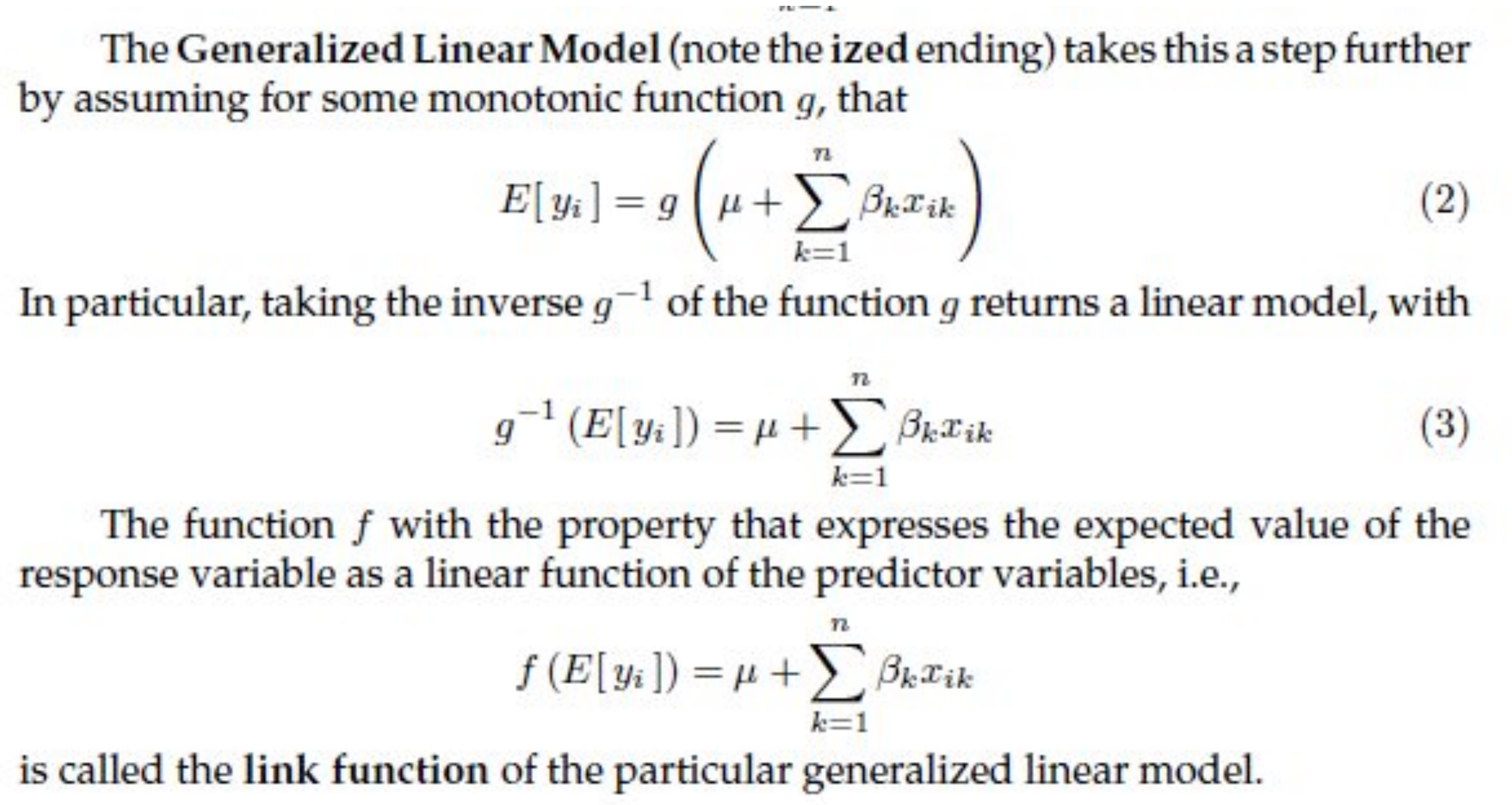
Different statistical models

- GLM = general linear model
  - OLS ordinary least squares: $e \sim MVN(0,cI)$
  - GLS generalized least squares: $e \sim MVN(0,R)$
- Mixed models
  - Both fixed and random effects (beyond the residual)
- Mixture models
  - A weighted mixture of distributions
- Generalized 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

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

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

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

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