Lecture 2: Linear Models

Bruce Walsh lecture notes
SISG -Mixed Model Course
version 28 June 2012

Quick Review of the Major Points

The general linear model can be written as

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

- **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** β

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

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

Ordinary least squares (OLS)

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

Generalized least squares (GLS)

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

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 $\text{Var}(\mathbf{e}) = \sigma_e^2 \mathbf{I}$ -- OLS
  - All others, i.e., $\text{Var}(\mathbf{e}) = \mathbf{R}$ -- 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, $\mathbf{e}^T\mathbf{e} = \Sigma\, e_i^2$

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 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**

The Sire Model in Matrix Form

The model here is $y_{ij} = \mu + s_i + e_{ij}$
Consider three sires. Sire 1 has 2 offspring, sire 2 has one and sire 3 has three. The GLM form is

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{31} \\ y_{32} \\ y_{33} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ s_1 \\ s_2 \\ s_3 \end{pmatrix}, \quad \text{and} \quad \mathbf{e} = \begin{pmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{31} \\ e_{32} \\ e_{33} \end{pmatrix}$$

Still need to specify covariance structure of the residuals. For example, could be common-family effects that make some correlated.

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 $\det(\mathbf{X}^T\mathbf{X}) = 0$, 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 $\mathrm{Var}(\beta) = k\,(X^TX)^{-1}$
- 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

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

Predicted value of the y's

Ordinary Least Squares (OLS)

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

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

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

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

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

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

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

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

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
- $Y_i = \beta_1 x_i + \beta_2 x_i^2 + \beta_3 s_i x_i^2$
- s_i is an indicator (0/1) variable for the sex (0=male, 1=female).
  - Male slope = β_2,
  - Female slope = $\beta_2 + \beta_3$

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 and predicting random effects.

"Mixed" models contain both fixed and random factors

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

Key: need to specify covariance structures for MM

Example: Sire model

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

Here μ is a fixed effect, e a random effect

Is the sire effect s fixed or random ?

It depends. If we have (say) 10 sires, if we are ONLY interested in the values of these particular 10 sires and don't care to make any other inferences about the population from which the sires are drawn, then we can treat them as fixed effects. In the case, the model is fully specified the covariance structure for the residuals. Thus, we need to estimate μ, s_1 to s_{10} and σ^2_e, and we write the model as $y_{ij} = \mu + s_i + e_{ij}$, $\sigma^2(e) = \sigma^2_e I$

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

Conversely, if we are not only interested in these 10 particular sires but also wish to make some inference about the population from which they were drawn (such as the additive variance, since $\sigma^2_A = 4\sigma^2_s$,), then the s_i are random effects. In this case we wish to estimate μ and the variances σ^2_s and σ^2_e. Since $2s_i$ also estimates (or predicts) the breeding value for sire i, we also wish to estimate (predict) these as well. Under a Random-effects interpretation, we write the model as $y_{ij} = \mu + s_i + e_{ij}$, $\sigma^2(e) = \sigma^2_e I$, $\sigma^2(s) = \sigma^2_A A$

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

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

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

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

Example

One common setting is where the residuals are uncorrelated but heteroscedastic, with $\sigma^2(e_i) = \sigma^2_e/w_i$.

For example, sample i is the mean value of n_i individuals, with $\sigma^2(e_i) = \sigma_e{}^2/n_i$. Here $w_i = n_i$

$$\mathbf{R} = \mathrm{Diag}(w_1^{-1}, w_2^{-1}, \ldots, w_n^{-1})$$

Consider the model $y_i = \alpha + \beta x_i$

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

Here

$$\mathbf{R}^{-1} = \mathrm{Diag}(w_1, w_2, \ldots, w_n)$$

giving

$$\mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} = w \begin{pmatrix} \overline{y}_w \\ \overline{xy}_w \end{pmatrix} \qquad \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} = w \begin{pmatrix} 1 & \overline{x}_w \\ \overline{x}_w & \overline{x^2}_w \end{pmatrix}$$

where

$$w = \sum_{i=1}^n w_i, \quad \overline{x}_w = \sum_{i=1}^n \frac{w_i x_i}{w}, \quad \overline{x^2}_w = \sum_{i=1}^n \frac{w_i x_i^2}{w}$$

$$\overline{y}_w = \sum_{i=1}^n \frac{w_i y_i}{w}, \quad \overline{xy}_w = \sum_{i=1}^n \frac{w_i x_i y_i}{w}$$

This gives the GLS estimators of α and β as

$$a = \overline{y}_w - b\overline{x}_w$$

$$b = \frac{\overline{xy}_w - \overline{x}_w \overline{y}_w}{\overline{x^2}_w - \overline{x}_w^2}$$

Likewise, the resulting variances and covariance for these estimators is

$$\sigma^2(a) = \frac{\sigma_e^2 \cdot \overline{x^2}_w}{w\,(\overline{x^2}_w - \overline{x}_w^2)}$$

$$\sigma^2(b) = \frac{\sigma_e^2}{w\,(\overline{x^2}_w - \overline{x}_w^2)}$$

$$\sigma(a,b) = \frac{-\sigma_e^2 \overline{x}_w}{w\,(\overline{x^2}_w - \overline{x}_w^2)}$$

Chi-square and F distributions

Let $U_i \sim N(0,1)$, i.e., a unit normal

The sum $U_1^2 + U_2^2 + \cdots + U_k^2$ is a chi-square random variable with k degrees of freedom

Under appropriate normality assumptions, the sums of squares that appear in linear models are also chi-square distributed. In particular,

$$\sum_{i=1}^n (x_i - \bar{x})^2 \sim \chi^2_{n-1}$$

The ratio of two chi-squares is an **F distribution**

In particular, an F distribution with k numerator degrees of freedom, and n denominator degrees of freedom is given by

$$\frac{\chi_k^2/k}{\chi_n^2/n} \sim F_{k,n}$$

Thus, F distributions frequently arise in tests of linear models, as these usually involve ratios of sums of squares.

Sums of Squares in linear models

The total sums of squares SST of a linear model can be written as the sum of the error (or residual) sum of squares and the model (or regression) sum of squares

$$SS_T = SS_M + SS_E$$

$$\sum (y_i - \overline{y})^2 \quad \sum (\widehat{y}_i - \overline{y})^2 \quad \sum (y_i - \widehat{y}_i)^2$$

r^2, the coefficient of determination, is the fraction of variation accounted for by the model

$$r^2 = \frac{SS_M}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

Sums of Squares are quadratic products

$$SS_T = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - \overline{y}^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n} y_i\right)^2$$

We can write this as a quadratic product as

$$SS_T = \mathbf{y}^T\mathbf{y} - \frac{1}{n}\mathbf{y}^T\mathbf{J}\mathbf{y} = \mathbf{y}^T\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{y}$$

Where **J** is a matrix all of whose elements are 1's

$$SS_E = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} \widehat{e}_i^2$$

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

$$SS_M = SS_T - SS_E = \mathbf{y}^T\left(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T - \frac{1}{n}\mathbf{J}\right)\mathbf{y}$$

Expected value of sums of squares

- In ANOVA tables, the E(MS), or expected value of the Mean Squares (scaled SS or Sum of Squares), often appears
- This directly follows from the quadratic product. If $E(\mathbf{x}) = \mu$, $Var(\mathbf{x}) = \mathbf{V}$, then
  - $E(\mathbf{x}^T\mathbf{A}\mathbf{x}) = tr(\mathbf{A}\mathbf{V}) + \mu^T\mathbf{A}\mu$

Hypothesis testing

Provided the residual errors in the model are MVN, then for a model with n observations and p estimated parameters,

$$\frac{\mathrm{SS}_E}{\sigma_e^2} \sim \chi^2_{n-p}$$

Consider the comparison of a full (p parameters) and reduced (q < p) models, where SSE_r = error SS for reduced model, SSE_f = error SS for full model

$$\left(\frac{\mathrm{SS}_{E_r} - \mathrm{SS}_{E_f}}{p-q}\right) \Big/ \left(\frac{\mathrm{SS}_{E_f}}{n-p}\right) = \left(\frac{n-p}{p-q}\right)\left(\frac{\mathrm{SS}_{E_r}}{\mathrm{SS}_{E_f}} - 1\right)$$

The difference in the error sum of squares for the full and reduced model provided a test for whether the model fit is the same

This ratio follows an $F_{p-q,n-p}$ distribution

Does our model account for a significant fraction of the variation?

Here the reduced model is just $y_i = u + e_i$

In this case, the error sum of squares for the reduced model is just the total sum of squares and the F test ratio becomes

$$\left(\frac{n-p}{p-1}\right)\left(\frac{\mathrm{SS}_T}{\mathrm{SS}_{E_f}} - 1\right) = \left(\frac{n-p}{p-1}\right)\left(\frac{r^2}{1-r^2}\right)$$

This ratio follows an $F_{p-1,n-p}$ distribution

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 $\boldsymbol{\beta}$	$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$	$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$
$\mathrm{Var}(\widehat{\boldsymbol{\beta}})$	$(\mathbf{X}^T\mathbf{X})^{-1}\sigma_e^2$	$(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$
Predicted values, $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$	$\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$	$\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$
$\mathrm{Var}(\widehat{\mathbf{y}})$	$\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma_e^2$	$\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T$

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
- 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

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

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

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

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.,

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

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

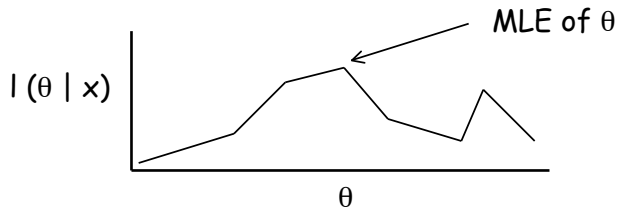
Typically assume non-normal distribution for residuals, e.g., Poisson, binomial, gamma, etc

Different methods of analysis

- Parameters of these various models can be estimated in a number of frameworks
- Method of moments
  - Very little assumptions about the underlying distribution. Typically, the mean of some statistic has an expected value of the parameter
  - OLS & GLS examples. We only need the assumption on the covariance structure of the residuals and finite moments.
  - While estimation does not require distribution assumptions, confidence intervals and hypothesis testing does
- Distribution-based estimation
  - The explicit form of the distribution used

Distribution-based estimation

- Maximum likelihood estimation
  - MLE
  - REML
  - More in Lynch & Walsh Appendix 3
- Bayesian
  - Marginal posteriors
  - Conjugating priors
  - MCMC/Gibbs sampling
  - More in Walsh & Lynch Appendices 2,3



This is formalize by looking at the log-likelihood surface, L = ln [l (θ | x)]. Since ln is a monotonic function, the value of θ that maximizes l also maximizes L

The curvature of the likelihood surface in the neighborhood of the MLE informs us as to the precision of the estimator A narrow peak = high precision. A board peak = lower precision

$$\text{Var(MLE)} = -1 \Big/ \frac{\partial^2 L(\mu|\mathbf{z})}{\partial \mu^2}$$

The larger the curvature, the smaller the variance

Maximum Likelihood

$p(x_1,\ldots, x_n \mid \theta)$ = density of the observed data $(x_1,\ldots, x_n)$ given the (unknown) distribution parameter(s) θ

Fisher suggested the method of maximum likelihood --- given the data $(x_1,\ldots, x_n)$ find the value(s) of θ that maximize $p(x_1,\ldots, x_n \mid \theta)$

We usually express $p(x_1,\ldots, x_n \mid \theta)$ as a likelihood function $l(\theta \mid x_1,\ldots, x_n)$ to remind us that it is dependent on the observed data

The Maximum Likelihood Estimator (MLE) of θ are the value(s) that maximize the likelihood function l given the observed data $x_1,\ldots, x_n$.

Likelihood Ratio tests

Hypothesis testing in the ML frameworks occurs through likelihood-ratio (LR) tests

$$LR = 2\ln\left(\frac{\ell(\widehat{\Theta}_r \mid \mathbf{z})}{\ell(\widehat{\Theta} \mid \mathbf{z})}\right) = 2\left[L(\widehat{\Theta}_r \mid \mathbf{z}) - L(\widehat{\Theta} \mid \mathbf{z})\right]$$

θ_r is the MLE under the restricted conditions (some parameters specified, e.g., var =1)

Θ_r is the MLE under the unrestricted conditions (no parameters specified)

For large sample sizes (generally) LR approaches a Chi-square distribution with r df (r = number of parameters assigned fixed values under null)

Likelihoods for GLMs

Under assumption of MVN, x ~ MVN(β ,**V**), the likelihood function becomes

$$L(\beta,\mathbf{V} \mid \mathbf{x}) = (2\pi)^{-n/2} |\mathbf{V}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\beta)^T \mathbf{V}^{-1} (\mathbf{x}-\beta)\right]$$

Variance components (e.g., σ^2_A, σ^2_e, etc.) are included in **V**

REML = restricted maximum likelihood. Method of choice for variance components, as it maximizes that part of the likelihood function that is independent of the fixed effects, β.

Bayesian Statistics

An extension of likelihood is Bayesian statistics

Instead of simply estimating a point estimate (e.g.., the MLE), the goal is the estimate the entire distribution for the unknown parameter θ given the data x

$$p(\theta \mid x) = C * l(x \mid \theta)\, p(\theta)$$

$p(\theta \mid x)$ is the posterior distribution for θ given the data x

$l(x \mid \theta)$ is just the likelihood function

$p(\theta)$ is the prior distribution on θ.

Bayesian Statistics

Why Bayesian?

- Exact for any sample size
- Marginal posteriors
- Efficient use of any prior information
- MCMC (such as Gibbs sampling) methods

Priors quantify the strength of any prior information. Often these are taken to be diffuse (with a high variance), so prior weights on θ spread over a wide range of possible values.

Marginal posteriors

- Often times were are interested in a particular set of parameters (say some subset of the fixed effects). However, we also have to estimate all of the other parameters.
- How do uncertainties in these nuisance parameters factor into the uncertainty in the parameters of interest?
- A Bayesian marginal posterior takes this into account by integrating the full posterior over the nuisance parameters
- While this sound complicated, easy to do with MCMC.

Conjugating priors

For any particular likelihood, we can often find a conjugating prior, such that the product of the likelihood and the prior returns a known distribution.

Example: For the mean μ in a normal, taking the prior on the mean to also be normal returns a posterior for μ that is normal.

Example: For the variance σ^2 in a normal, taking the prior on the variance to an inverse chi-square distribution returns a posterior for σ^2 that is also an inverse chi-square (details in WL Appendix 2).

A normal prior on the mean with mean μ_0 and variance σ_0^2 (larger σ_0^2, more diffuse the prior

$$p(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)$$

If the likelihood for the mean is a normal distribution, the resulting posterior is also normal, wit

$$\sigma_*^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1} \quad \text{and} \quad \mu_* = \sigma_*^2\left(\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}\right)$$

Note that if σ_0^2 is large, the mean of the posterior is very close to the sample mean.

If x follows a Chi-square distribution, then 1/x follows an inverse chi-square distribution.

The **scaled inverse chi-square distribution** is more typically used, where

$$p(x \mid n) \propto x^{-(n/2+1)} e^{-\sigma_0^2/(2x)}$$

The scaled inverse chi-square has two parameters, allowing more control over the mean and variance of the prior

Likelihood	Conjugate prior
Normal	
μ unknown, σ^2 known	Normal
μ known, σ^2 unknown	Inverse Chi-Square
Multivariate Normal	
$\boldsymbol{\mu}$ unknown, $\mathbf{V}$ known	Multivariate Normal
$\boldsymbol{\mu}$ known, $\mathbf{V}$ unknown	Inverse Wishart

MCMC

Analytic expressions for posteriors can be complicated, but the method of MCMC (Markov Chain Monte Carlo) is a general approach to simulating draws for just about any distribution (details in WL Appendix 3).

Generating several thousand such draws from the posterior returns an empirical distribution that we can use.

For example, we can compute a 95% credible interval, the region of the distribution that containing 95% of the probability.

Gibbs Sampling

- A very powerful version of MCMC is the Gibbs Sampler
- Assume we are sampling from a vector of parameters, but that the marginal distribution of each parameter is known
- For example, given a current value for all the fixed effects (but one, say β_1) and the variances, conditioning on these values the distribution of β_1 is a normal, whose parameters are now functions of the current values of the other parameters. A random draw is then generated from this distribution.
- Likewise, conditioning on all the fixed effects and all variances but one, the distribution of this variance is an inverse chi-square

When more than two variables are involved, the sampler is extended in the obvious fashion. In particular, the value of the kth variable is drawn from the distribution $p(\theta^{(k)} \mid \boldsymbol{\Theta}^{(-k)})$ where $\boldsymbol{\Theta}^{(-k)}$ denotes a vector containing all of the variables but k. Thus, during the ith iteration of the sample, to obtain the value of $\theta_i^{(k)}$ we draw from the distribution

$$\theta_i^{(k)} \sim p(\theta^{(k)} \mid \theta^{(1)} = \theta_i^{(1)}, \cdots, \theta^{(k-1)} = \theta_i^{(k-1)}, \theta^{(k+1)} = \theta_{i-1}^{(k+1)}, \cdots, \theta^{(n)} = \theta_{i-1}^{(n)})$$

For example, if there are four variables, (w, x, y, z), the sampler becomes

$$\begin{aligned} w_i &\sim p(w \mid x = x_{i-1}, y = y_{i-1}, z = z_{i-1}) \\ x_i &\sim p(x \mid w = w_i, y = y_{i-1}, z = z_{i-1}) \\ y_i &\sim p(y \mid w = w_i, x = x_i, z = z_{i-1}) \\ z_i &\sim p(z \mid w = w_i, x = x_i, y = y_i) \end{aligned}$$

This generates one cycle of the sampler. Using these new values, a second cycle is generated.

Full details in WL Appendix 3.