Theory of Linear Models

Steven Gilmour
King's College London

January – February 2020

Components of Variance

In many studies there are several nested levels of variation, e.g. schools and pupils.

It is natural to consider each of these as contributing a variance component, e.g. for group i, unit j, we have

$$Y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \delta_i + \epsilon_{ij},$$

where $\delta_i \sim N(0, \sigma_1^2)$, $\epsilon_{ij} \sim N(0, \sigma^2)$ and all r.v.s are independent.

This can be rewritten as the linear mixed model

$$\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}),$$

where

$$\mathbf{V} = \begin{bmatrix} \mathbf{U} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{U} & & \\ \vdots & & \ddots & \\ \mathbf{0} & & & \mathbf{U} \end{bmatrix}$$

and

$$\mathbf{U} = \begin{bmatrix} \sigma_1^2 + \sigma^2 & \sigma_1^2 & \cdots & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 + \sigma^2 & & \\ \vdots & & \ddots & \\ \sigma_1^2 & & & \sigma_1^2 + \sigma^2 \end{bmatrix}.$$

This generalizes in an obvious way to:

- unequal numbers of units in each group;
- any number of variance components;
- different unit variances in different groups;
- negative variance components.

If σ_1^2/σ^2 were known, $\boldsymbol{\beta}$ could be estimated by generalized least squares.

In practice, we usually estimate the variance components and plug the estimates in to a generalized least squares fit - empirical generalized least squares.

This underestimates the standard errors of $\hat{\boldsymbol{\beta}}$, but various corrections are available (Satterthwaite, Kenward-Roger).

The variance components can be estimated by maximum likelihood or, more commonly, residual maximum likelihood (REML) (or we can do a Bayesian analysis).

REML Estimation

REML uses an orthogonal transformation of $\mathbf{Y}$ of dimension $n - p$ and maximises this residual likelihood.

We use the transformation

$$\mathbf{Y}^* = (\mathbf{I} - \mathbf{H})\mathbf{Y},$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Then

$$\mathbf{Y}^* \sim N_n(\mathbf{0}, (\mathbf{I} - \mathbf{H})\mathbf{V}(\mathbf{I} - \mathbf{H}))$$

and we maximise the likelihood obtained from this, using a generalised inverse of $Var(\mathbf{Y}^*)$, to obtain the REML estimates of σ^2 and σ_1^2.

In the general linear model, REML gives S^2 as the estimator of σ^2.

Relaxing the Normality Assumption

We can obtain expressions for minimum norm quadratic unbiased estimators (MINQUEs) of variance components.

Except in simple cases, these depend on the unknown values of the variance components.

If we iterate (I-MINQUE), the estimates converge to the REML estimates.

Hence, REML estimators do not really depend on distributional assumptions.

A Warning

In complex structured data sets, REML (and other) estimates of variance components can turn out to be zero (or very close to zero). This can be because:

- true value is (very close to) zero; or
- the data do not provide enough information to estimate this variance component.

It is usually impossible to tell which of these is true.

The E-GLS estimator of fixed effects simply plugs in these estimates in order to estimate the fixed effects.

Standard errors of fixed effect estimators are calculated from $(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}$.

This acts as if the variance components are known to be zero.

There is no good solution to this (except fully Bayesian methods).

Generalized Linear Models

Linear models are useful for response variables taking values on $\mathbb{R}$.

Generalized linear models (GLMs) are used for other responses.

GLMs require an assumption that the responses are from a particular distribution from the exponential family, with location parameter μ depending on $\mathbf{x}$ through a link function $g(\cdot)$:

$$g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}$$

and constant dispersion parameter.

These are most commonly used with discrete data, the assumed distribution being Poisson or binomial, but occasionally the gamma distribution is used for continuous data on $\mathbb{R}^+$. This is an alternative to transforming Y.

GLMs have a number of properties which make them simple to work with:

- Maximum likelihood estimates can be obtained using iteratively reweighted least squares, which ensures convergence.
- Canonical link functions (e.g. $\log \lambda$ for Poisson, $\log\{\pi/(1-\pi)\}$ for binomial) can be used, for which $\mathbf{X}'\mathbf{Y}$ is a sufficient statistic.

Quasi-likelihood models retain the same mean-variance relationship as the distributions used in GLMs, but drop the distributional assumption and allow for an extra dispersion parameter.

For example, instead of assuming $Y \sim \text{Poisson}(\lambda)$, which implies $E(Y) = \lambda$ and $V(Y) = \lambda$, we assume $E(Y) = \lambda$ and $V(Y) = \sigma^2\lambda$, with the same link function, but without assuming a specific distribution.

This is a kind of semi-parametric model.

Mixed Models for Non-Normal Responses

Additional random effects can be built in to GLMs in two ways. Both require distributional assumptions on the random effects, except at the bottom level, where we can use a quasi-likelihood model.

Generalised linear mixed models (GLMMs) use normally distributed random effects in the linear predictor, e.g.

$$g(\mu_{ij}|\delta_i) = \mathbf{x}_{ij}\boldsymbol{\beta} + \delta_i,$$

where μ_{ij} is the expected response from the jth unit in group i and $\boldsymbol{\delta} \sim N_n(\mathbf{0}, \sigma_b^2\mathbf{I})$.

GLMMs are usually fitted by maximum likelihood, using Gauss-Hermite quadrature.

Other Models for Random Effects

Hierarchical generalised linear models (HGLMs) relax the normality assumption for the ransom effects. Their distribution is completely general, but things simplify if a canonical distribution is used, e.g.

$$Y_{ij}|\lambda_{ij} \sim \text{Poisson}(\lambda_{ij}),$$

with

$$\lambda_i j \sim \text{Gamma}(\mu_{ij}, \theta)$$

and $\mu_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta}$.

These can be fitted using the h-likelihood, though this is controversial.

Bayesian methods are completely flexible, but computationally intensive to fit.

Proportional Hazards Models

For time-to-event data $\mathbf{T}$, it is often appropriate to assume the model

$$T_i \sim \text{Weibull}(\alpha_i, \eta),$$

where $\log \alpha_i = \mathbf{x}_i' \boldsymbol{\beta}$.

This is not a GLM, since the Weibull distribution is not a member of the exponential family.

The hazard function for this distribution is

$$\begin{aligned} h(t) = \frac{f(t)}{1 - F(t)} &= \eta \alpha_i^{-\eta} t^{\eta - 1} \\ &= \eta e^{-\eta \mathbf{x}_i' \boldsymbol{\beta}} t^{\eta - 1} \\ &= h_0(t) e^{\mathbf{x}_i' \boldsymbol{\beta}}, \end{aligned}$$

where $h_0(t)$ is the baseline hazard for an observation with $\mathbf{x} = \mathbf{0}$.

The Weibull regression model has the property of proportional hazards.

Very often, we use Cox's semi-parametric proportional hazards model, which assumes

$$h(t) = h_0(t)e^{\mathbf{x}_i'\boldsymbol{\beta}},$$

but does not make any distributional assumption.

As usual, proportional hazards is just an assumption and can be relaxed. We can also include random effects (frailty) in these models.

Semi-Parametric Models

Semi-parametric regression models assume that $Y \sim N(f(\mathbf{x}), \sigma^2)$, but avoid assume a specific functional form for $f(\mathbf{x})$.

Instead, data-driven smoothers are used to approximate the unknown function. Typical examples are smoothed local polynomials (splines), which try to follow the data, while penalising "roughness".

These can be extended to other distributions using generalized additive models.

Opinion: These models are most useful for nuisance effects, as they do not allow mechanistic understanding to be developed.

Again there is no problem in including random effects of any required complexity, though fitting the models can be a computational challenge.

Nonlinear Models

The term "nonlinear models" is completely general, but is most often associated with models of the form $E(Y_i) = f(\mathbf{x}_i; \boldsymbol{\theta})$ and $V(\mathbf{Y}) = \sigma^2 \mathbf{I}$.

The parameters are usually estimated by nonlinear least squares (NLLS), i.e. we choose $\boldsymbol{\theta}$ to minimise

$$\sum_{i=1}^{n} \{Y_i - f(\mathbf{x}_i; \boldsymbol{\theta})\}^2.$$

Inference is usually carried out by assuming normality, in which case the NLLS estimates are equivalent to the MLEs, and invoking the asymptotic properties of MLEs.

Nonlinear least squares is a fundamentally difficult numerical problem: Convergence to the global optimum is not guaranteed and often fails in practice.

Partial linear algorithms can help when there are separable linear parameters, i.e. the model can be written

$$E(Y) = \sum_{j=1}^{q} \beta_j f_j(\mathbf{x}; \boldsymbol{\theta}).$$

This reduces the numerical search by q dimensions.

Also, asymptotic inferences can be very poor approximations.

Bayesian methods avoid *these* computational problems, but might create others.

We can also use nonlinear mixed models.

Empirical Nonlinear Models

Models based on (artificial) neural networks are often very successful in practice.

They can be viewed as very complex, but purely empirical, nonlinear regression (or classification) models, with
$E(Y_i|\mathbf{z}_i) = f_z(\mathbf{z}_i;\boldsymbol{\theta}_z)$, $E(Z_i|\mathbf{a}_i) = f_a(\mathbf{a}_i;\boldsymbol{\theta}_a)$, $\cdots$,
$E(W_i|\mathbf{x}_i) = f_x(\mathbf{x}_i;\boldsymbol{\theta}_x)$.

The functional form and number of dimensions used in each "layer" are arbitrary and usually chosen purely empirically.

Implicitly, we usually assume $V(\mathbf{Y}) = \sigma^2\mathbf{I}$.

Conceptually, there is no difficulty in including random effects of the appropriate structure; the only restriction is computational.

Statistical Modelling and Machine Learning

Machine learning methods are regression modelling methods, typically with the following features:

- the fixed effects model is complex, purely empirical and often implicit;
- the random effects model is implicitly iid $N(0, \sigma^2)$;
- there is an emphasis on prediction;
- the model is used as a black box;
- there is an emphasis on big data sets and/or real-time predictions.

Statistical Modelling versus Machine Learning?

It is easy to see that we can do better prediction by more realistically modelling the random effects structure, we can gain more understanding of the system and be better able to assess the quality of predictions by doing inference,

Statistical Modelling versus Machine Learning?

It is easy to see that we can do better prediction by more realistically modelling the random effects structure, we can gain more understanding of the system and be better able to assess the quality of predictions by doing inference, if we have the computational power to do this in the time available before a prediction is required.

If there is insufficient computational power, can we do better by making the fixed effects structure simpler and the random effects structure more complex? I believe this is an open question in most cases.