Theory of Linear Models

Steven Gilmour
King's College London

January – February 2020

Analysis of Designed Experiments

Theory of analysis of data from designed experiments is based on the randomisation of the experiment determining the appropriate analysis of variance and linear model.

Designed experiments and linear models are inextricably linked, but sometimes more advanced modelling is appropriate.

Often experimenters:

- do over-simple design and/or analysis; or
- ignore the design when analysing the data and/or ignore the appropriate analysis when designing the experiment.

Here, we aim to link several types of model, more advanced than linear models, to designed experiments.

The emphasis is on the link between the *structure* of the design and the precise formulation of the model.

Similar ideas apply to data from structured sample surveys, or indeed any data where there is a clear structure, nested, crossed or a combination of these.

Randomisation Analysis of Experimental Data

Planning an experiment:

1. objectives lead to treatments;
2. experimental units identified;
3. unit structure defined;
4. the responses are identified;
5. restrictions on allocation of treatments to units identified;
6. treatments allocated to units to get efficient design.

Randomisation analysis requires only the assumption that

$$\begin{pmatrix} \text{Response in unit} \\ i \text{ when treatment} \\ r \text{ is applied} \end{pmatrix} =$$

$$\begin{pmatrix} \text{Quantity} \\ \text{depending only} \\ \text{on the unit} \end{pmatrix} + \begin{pmatrix} \text{Quantity} \\ \text{depending only} \\ \text{on the treatment} \end{pmatrix}$$

Consider a field experiment in 20 plots to compare the yields of 4 varieties of crop each replicated 5 times. Assume that

$$y_{i(r)} = m + e_i + t_r,$$

where $y_{i(r)}$ is the yield from plot i if variety r is applied, where $\sum_{i=1}^{20} e_i = \sum_{r=1}^{4} t_r = 0$.

Unrandomised design:

Plot	Variety	Plot	Variety
1	1	11	3
2	1	12	3
3	1	13	3
4	1	14	3
5	1	15	3
6	2	16	4
7	2	17	4
8	2	18	4
9	2	19	4
10	2	20	4

Randomly allocate plot labels to plots.

The above deterministic model now becomes stochastic over the population of possible randomisations, i.e.

$$Y_{i(r)} = m + \epsilon_i + t_r,$$

where $\epsilon_i = e_j$ for some j determined by the outcome of randomisation.

Note that ϵ_i has expectation 0 and constant variance.

This is where the usual linear model comes from. No assumption about a population from which the units are sampled has been made.

Other orthogonal block structures can be developed by rewriting the unit effect e_i in an appropriate way, e.g. with r rows and c columns we write

$$y_{ij(r)} = m + r_i + c_j + e_{ij} + t_r,$$

which under randomisation becomes

$$Y_{ij(r)} = m + \rho_i + \gamma_j + \epsilon_{ij} + t_r.$$

Appropriate randomisations ensure:

- least squares estimators are BLUEs of all treatment contrasts, e.g. factorial contrasts, linear and quadratic contrasts;
- MS_{Res} in each stratum is BUE of the variance for that stratum.

Other Forms of Analysis

The obvious alternative is to assume that the units are a sample from a (possibly structured) population of potential units.

This allows inferences to be made about the population of potential units, not just those used in the experiment, but is reliant on the additional assumptions about the nature of the population.

Then, for example, the randomised block model arises if we assume the units in different blocks are random samples from different populations. If the block populations are assumed to be a sample from a super-population, we have random block effects, as in the randomisation analysis.

If we randomise the experiment in a way that corresponds to the population structure, the two analyses are identical.

This is obviously a good idea!

Then only the generalisation from the units used to the population depends on the assumptions made about the population. The analysis will be valid in any case.

Pure randomisation based analysis is frequentist, but it provides a natural baseline for a model based analysis, whether likelihood based or Bayesian.

Even the purist Bayesian viewpoint does not *object* to randomisation and should not object to including blocking factors defined by the randomisation in the model.

Doing this allows us to separate:

- conclusions obtained from the randomisation analysis (which are robust to assumptions);
- conclusions which depend on the model assumptions;
- conclusions which depend on a particular prior distribution.

Nearly Saturated Factorial Treatment Structures

To investigate the effects of temperature (coded as X_1) and pressure (coded as X_2) on the yield of a reaction, the first experiment might use the design:

Treat	X_1	X_2
1	-1	-1
2	-1	1
3	1	-1
4	1	1
5	0	0
5	0	0
5	0	0
5	0	0

Obtain the following analysis of variance:

Source	df
Temperature_L	1
Pressure_L	1
$\text{Temp}_L \times \text{Press}_L$	1
Residual:	4:
Lack of fit	1
Pure error	3
Total	7

From the randomisation viewpoint, we obtain 4 orthogonal treatment contrasts:

Source	df
Treatments:	4:
Temperature_L	1
Pressure_L	1
$\text{Temp}_L \times \text{Press}_L$	1
Lack of fit	1
Residual	3
Total	7

This makes clearer the meaning of "pure error" - it is just the usual unbiased estimator of σ^2.

If lack of fit, or the interaction, are close to zero, should we replace the unbiased estimator of σ^2 with a biased one including these terms in the residual?

No, unless there is some reason why we must get as good an estimate from this small experiment as possible. Not usually the case.

Block Design

Block I			Block II		
Treat	X_1	X_2	Treat	X_1	X_2
1	-1	-1	1	-1	-1
2	-1	1	2	-1	1
3	1	-1	3	1	-1
4	1	1	4	1	1
5	0	0	5	0	0
5	0	0	5	0	0

How many df for pure error?

Only the randomisation analysis gives a sensible answer.

Source	df
Blocks	1
Treatments:	4:
Temperature_L	1
Pressure_L	1
$\text{Temp}_L \times \text{Press}_L$	1
Lack of fit	1
Residual:	6:
"Lack of fit"	4
"Pure error"	2
Total	11

Saturated Structures

To study the effects of catalyst (two types), amount of chemical A (low/high), amount of chemical B (low/high), stirring (yes/no) and shaking (yes/no) on a chemical system, use a single replicate of 2^5 factorial treatment combinations completely randomised.

Source	df
Catalyst	1
Chemical A	1
Chemical B	1
Stirring	1
Shaking	1
2-factor interactions	10
3-factor interactions	10
4-factor interactions	5
5-factor interaction	1
Residual	0
Total	31

Get BLUEs of all factorial effects, but no estimator of σ^2 with which to carry out inference.

Possible solutions:

1. Do nothing! Exploratory analysis is all we need, e.g. Normal plot of estimated effects.
2. Estimate σ^2 from the small effects in the Normal plot. Biased estimator.
3. Assume *a priori* that high order interactions will be zero. Strong assumption and several small, but non-zero, effects can cause unquantifiable bias. Given this assumption, 2 replicates of a half-fraction would be better.

4. Use a prior estimate of σ^2. Strong assumption.
5. Use a prior distribution for σ^2. Experiment provides no information about σ^2, so still a strong assumption.
6. Perform a fully Bayesian analysis, with priors on each effect and updating all priors using Bayes' Theorem.

I recommend 1 or 6, depending on what conclusions we want to draw.

Supersaturated Structures

Study more factors than there are experimental units, e.g. crash testing cars.

No sensible randomisation analysis, perhaps fully Bayesian analysis is the only sensible one.

Quantitative Treatment Structures

Consider an experiment in enzyme kinetics with 3 replicates at each substrate concentration 10, 20, 40, 80, 160, completely randomised.

Biochemical theory implies

$$E(Y_{ij}) = \frac{\theta_0 x_i}{\theta_1 + x_i}.$$

Natural to fit this by nonlinear least squares - gives estimates of θ_0, θ_1 and σ^2.

This is not an unbiased estimator of σ^2. Get this from the randomisation analysis. Difference represents lack of fit.

Source	df
Treatments:	4:
Michaelis-Menten model	1
Lack of fit	3
Residual	10
Total	14

The Michaelis-Menten component in the analysis of variance does not correspond to a *linear* contrast, but the interpretation is the same.

If we use a randomised complete block design, we simply add a random effect for blocks as usual.

Source	df
Blocks	2
Treatments:	4:
Michaelis-Menten model	1
Lack of fit	3
Residual	8
Total	14

The error structure of NLLS has been questioned and instead a transform-both-sides model suggested:

$$Y_{ij}^{(\lambda)} = \left(\frac{\theta_0 x_i}{\theta_1 + x_i} \right)^{(\lambda)} + \epsilon_{ij},$$

where

$$Y^{(\lambda)} = \begin{cases} \frac{Y^\lambda - 1}{\lambda}, & \lambda \neq 0; \\ \log Y, & \lambda = 0. \end{cases}$$

This is a good idea, but can be difficult to fit in practice.

Randomisation theory allows a two-stage analysis which *greatly* simplifies the computation.

The assumption now is that unit and treatment effects are additive *on some transformed scale*, i.e.

$$y_{i(r)}^{(\lambda)} = m + e_i + t_r.$$

We fit the model

$$Y_{i(r)}^{(\lambda)} = m + \epsilon_i + t_r,$$

i.e. estimate a Box-Cox transformation for a completely randomised design model.

Now fixing $\hat{\lambda}$, use NLLS to fit

$$Y_{ij}^{(\hat{\lambda})} = \left(\frac{\theta_0 x_i}{\theta_1 + x_i} \right)^{(\hat{\lambda})} + \epsilon_{ij},$$

adjusting the residual term by one df.

Models for Complex Unit Effects

Spatial analysis of field experiments and modelling time trends in laboratory experiments have become fashionable.

The reason is an alleged increase in precision.

These methods involve replacing the unit effect, e_i, in our basic model with something more complex, e.g. an AR(1)$\times$AR(1) process in field experiments.

Block effects should be included in the model, so that we can separate conclusions that depend on the modelling assumptions from those that do not.

The model implied by the randomisation will often be adequate, so that no other strong assumptions are needed.

When the stronger assumptions lead to great increases in efficiency, it is often a sign that the experiment has been badly designed, e.g. with blocks that are too large.

Rescuing badly designed experiments seems to be the main purpose of these methods. However, the "rescue" is so dependent on assumptions that it is questionable whether these should still be called experiments.

Saturated Structures in Sequential Runs

A small experiment to study the effects of moisture content (coded as X_1), addition of enzyme (coded as X_2), mixing speed (coded as X_3) and baking temperature (coded as X_4) on properties of bread.

A half-fraction, units are run one after another.

Time	X_1	X_2	X_3	X_4
1	-1	-1	1	1
2	1	-1	-1	1
3	1	1	1	1
4	1	1	-1	-1
5	-1	1	-1	1
6	-1	-1	-1	-1
7	-1	1	1	-1
8	1	-1	1	-1

A linear time trend is expected, but the experiment is too small to sensibly use blocks of size 2.

We could completely randomise, then model unit effects as

$$e_i = \beta_1 v_i + \epsilon_i,$$

where v_i is the centred time of run i.

Since the full treatment model is inestimable, we might assume all interactions are zero and estimate the main effects.

This analysis is typical, but relies heavily on strong assumptions.

Better to either make neither of these assumptions or to do a fully Bayesian analysis.

Assume full factorial treatment model.

Independent priors for each factorial effect, e.g. $\beta_r \sim N(0, 4\sigma^2)$, $\beta_{rs} \sim N(0, 0.25\sigma^2)$, $\beta_{rst} \sim N(0, 0.01\sigma^2)$ and $\beta_{rstu} \sim N(0, 0.0001\sigma^2)$.

Assume general unit model, parameterised as a polynomial, i.e. $e_i = \beta_1 v_i + \beta_2 v_i^2 + \cdots + \beta_7 v_i^7 + \epsilon_i$.

Independent priors for each polynomial time effect, e.g. $\beta_1 \sim N(0, \sigma^2)$, $\beta_2 \sim N(0, 0.01\sigma^2)$, etc.

Prior for σ^2: scaled inverse-Gamma.

Typically will give a similar analysis to that described above, but with more uncertainty implied by the priors.

Discrete Data

Textbooks almost always present designed experiments in the context of a single, univariate, continuous response measured from each experimental unit. This is rare in practice. Can we deal with other situations?

Response is number of infected plants out of 30.

Given the possible outcomes, the original model,

$$y_{i(r)} = m + e_i + t_r,$$

makes no sense.

Instead a reasonable assumption seems to be

$$Y_{i(r)} \sim Binom(30, \pi_{i(r)}),$$

where $\pi_{i(r)}$ depends on the plot and the treatment.

We have assumed a distribution, but we can apply the ideas of randomisation based analysis to the unobservable linear predictor, e.g. assume

$$\log\left(\frac{\pi_{i(r)}}{1-\pi_{i(r)}}\right) = \eta_{i(r)} = m + e_i + t_r.$$

Under complete randomisation, this becomes

$$\eta_{i(r)} = m + \epsilon_i + t_r,$$

where ϵ_i is a random effect, as before.

If we approximate ϵ_i as being Normally distributed, we have a generalised linear mixed model (GLMM).

We can include block effects in the usual way.

Thus GLMs have no place in the analysis of data from designed experiments.

Longitudinal Data

An experiment to compare two growth hormones for cattle.

The experimental unit is the animal, but we have repeated measurements on each unit.

As with discrete data, we can first model the responses from each experimental unit, e.g. if y_{ij} is the jth measurement from animal i, assume

$$Y_{ij} = \beta_{0i} + \beta_{1i} v_{ij} + \epsilon_{ij},$$

where $E(\epsilon_{ij}) = 0$ and

$$Var(\boldsymbol{\epsilon}_i) = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}.$$

Combine this model with the usual randomisation model by assuming that the unobserved β_{0i} and the unobserved β_{1i} are additive in terms of unit and treatment effects.

Under randomisation we get

$$Y_{ij(r)} = m_0 + \alpha_i + t_{0r} + (m_1 + \gamma_i + t_{1r})v_{ij} + \epsilon_{ij},$$

where α_i and γ_i are random effects. This is the random slopes model.

Could apply to repeated measurements in space, rather than time.

Multivariate Responses

In many experiments there is more than one response variable. Usually, we analyse each one separately.

Consider an experiment to compare the effects of several factors on the baking pastry dough. Two responses are the cross-sectional expansion index and the longitudinal expansion index.

The multivariate analysis of variance model is

$$\begin{bmatrix} y_{1i(r)} \\ y_{2i(r)} \end{bmatrix} = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} + \begin{bmatrix} e_{1i} \\ e_{2i} \end{bmatrix} + \begin{bmatrix} t_{1r} \\ t_{2r} \end{bmatrix}.$$

Under randomisation, this becomes

$$\begin{bmatrix} y_{1i(r)} \\ y_{2i(r)} \end{bmatrix} = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{1i} \\ \epsilon_{2i} \end{bmatrix} + \begin{bmatrix} t_{1r} \\ t_{2r} \end{bmatrix},$$

where ϵ_{1i} and ϵ_{2i} are correlated because they are randomised to the same experimental unit.

This model implies that the treatments could have completely different effects on different responses.

A reasonable alternative model is

$$\begin{bmatrix} y_{1i(r)} \\ y_{2i(r)} \end{bmatrix} = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{1i} \\ \epsilon_{2i} \end{bmatrix} + \begin{bmatrix} t_{1r} \\ \phi t_{1r} \end{bmatrix}.$$

This can be fitted by NLLS.

Final Comments

The main message is that the model used should be determined initially by the design of the experiment, especially the randomisation.

If the basic theory of designed experiments had been worked out in the 1970s and 80s, instead of the 1920s and 30s, perhaps these methods would have been developed in the same way. The required computation that was not available before then.

There are many types of more complex models which are becoming available and the same ideas can be used with them.