MATH3014-6027 Design (and Analysis) of Experiments

Dave Woods

2022-01-23

Contents

Preface

These are draft lecture notes for the modules MATH3014 and MATH6027 Design (and Analysis) of Experiments at the University of Southampton for academic year 2021-22. They are very much work in progress.

Chapter 1

Motivation, introduction and revision

Definition 1.1. An **experiment** is the process through which data are collected to answer a scientific question (physical science, social science, actuarial science ...) by **deliberately** varying some features of the process under study in order to understand the impact of these changes on measureable responses.

In this course we consider only *intervention* experiments, in which some aspects of the process are under the experimenters' control. We do not consider *surveys* or *observational* studies.

Definition 1.2. Design of experiments is the topic in Statistics concerned with the selection of settings of controllable variables or factors in an experiment and their allocation to experimental units in order to maximise the effectiveness of the experiment at achieving its aim.

People have been designing experiments for as long as they have been exploring the natural world. Some notable milestones in the history of the design of experiments include:

- prior to the 20th century:
  - Francis Bacon (17th century; pioneer of the experimental methods)
  - James Lind (18th century; experiments to eliminate scurvy)
  - Charles Peirce (19th century; advocated randomised experiments and randomisation-based inference)
- 1920s: agriculture (particularly at the Rothamsted Agricultural Research Station)
- 1940s: clinical trials (Austin Bradford-Hill)
- 1950s: (manufacturing) industry (W. Edwards Deming; Genichi Taguchi)
- 1960s: psychology and economics (Vernon Smith)
- 1980s: in-silico (computer experiments)

- 2000s: online (A/B testing)

See Luca and Bazerman (2020) for further history, annecdotes and examples, especially from psychology and technology.

Figure 1.1 shows the Broadbalk agricultural field experiment at Rothamsted, one of the longest continuous running experiments in the world, which is testing the impact of different manures and fertilizers on the growth of winter wheat.

Figure 1.1: The Broadbalk experiment, Rothamsted (photograph taken 2016)

1.1 Motivation

Example 1.1. Consider an experiment to compare two treatments (e.g. drugs, diets, fertilisers, ...). We have N subjects (people, mice, plots of land, ...), each of which can be assigned one of the two treatments. A response (protein measurement, weight, yield, ...) is then measured.

Question: How many subjects should be assigned to each treatment to gain the most precise[1] inference about the difference in response from the two treatments?

Consider a linear statistical model[2] for the response (see MATH2010):

$$Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j\,, \qquad j = 1, \ldots, n\,, \tag{1.1}$$

[1] Smallest variance.

[2] In this course, we will almost always start with a statistical model which we wish to use to answer our scientific question.

where $\varepsilon_j \sim N(0, \sigma^2)$ are independent and identically distributed errors and β_0, β_1 are unknown constants (parameters).

Let[3]

$$x_j = \begin{cases} -1 & \text{if treatment 1 is applied to the } j\text{th subject} \\ +1 & \text{if treatment 2 is applied to the } j\text{th subject,} \end{cases}$$

for $j = 1, \ldots, n$.[4]

The difference in expected response from treatments 1 and 2 is

$$\begin{aligned} \mathrm{E}[Y_j \,|\, x_j = +1] - \mathrm{E}[Y_j \,|\, x_j = -1] &= \beta_0 + \beta_1 - \beta_0 + \beta_1 \\ &= 2\beta_1 \,. \end{aligned} \tag{1.2}$$

Therefore, we require the the most precise estimator of β_1 possible. That is, we wish to make the variance of our estimator of β_1 as small as possible.

Parameters β_0 and β_1 can be estimated using least squares (see MATH2010). For $Y_1, \ldots, Y_n$, we can write the model down in matrix form:

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} .$$

Or, by defining some notation:

$$Y = X\beta + \varepsilon \tag{1.3}$$

where

- Y - $n \times 1$ vector of responses;
- X - $n \times p$ model matrix;
- β - $p \times 1$ vector of parameters;
- ε - $n \times 1$ vector of errors.

The **least squares estimators**, $\hat{\beta}$, are chosen such that the quadratic form

$$(Y - X\beta)^{\mathrm{T}}(Y - X\beta)$$

is minimised (recall that $\mathrm{E}(\mathbf{Y}) = X\beta$). Therefore

$$\hat{\beta} = \mathrm{argmin}_{\beta}(Y^{\mathrm{T}}Y + \beta^{\mathrm{T}}X^{\mathrm{T}}X\beta - 2\beta^{\mathrm{T}}X^{\mathrm{T}}Y) \,.$$

[3] Other codings can be used: e.g. 0,1; see later in the module. It makes no difference for our current purpose.

[4] We will discuss the choice of *coding* -1, +1 later.

If we differentiate with respect to β[5],

$$\frac{\partial}{\partial \beta} = 2X^{\mathrm{T}}X\beta - 2X^{\mathrm{T}}Y\,,$$

and equate to 0, we get the estimators

$$\widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y\,. \tag{1.4}$$

These are the least squares estimators.

For Example 1.1,

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \qquad X^{\mathrm{T}}X = \begin{bmatrix} n & \sum x_j \\ \sum x_j & \sum x_j^2 \end{bmatrix},$$

$$(X^{\mathrm{T}}X)^{-1} = \frac{1}{n\sum x_j^2 - (\sum x_j)^2}\begin{bmatrix} \sum x_j^2 & -\sum x_j \\ -\sum x_j & n \end{bmatrix}, \qquad X^{\mathrm{T}}Y = \begin{bmatrix} \sum Y_j \\ \sum x_j Y_j \end{bmatrix}.$$

Then,

$$\begin{aligned} \widehat{\beta} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix} &= \frac{1}{n\sum x_j^2 - (\sum x_j)^2}\begin{bmatrix} \sum x_j^2 & -\sum x_j \\ -\sum x_j & n \end{bmatrix}\begin{bmatrix} \sum Y_j \\ \sum x_j Y_j \end{bmatrix} \\ &= \frac{1}{n\sum x_j^2 - (\sum x_j)^2}\begin{bmatrix} \sum Y_j \sum x_j^2 - \sum x_j \sum x_j Y_j \\ n\sum x_j Y_j - \sum x_j \sum Y_j \end{bmatrix}. \end{aligned} \tag{1.5}$$

We don't usually work through the algebra in such detail; the matrix form is often sufficient for theoretical and numerical calculations and software, e.g. `R`, can be used.

The precision of $\widehat{\beta}$ is measured via the variance-covariance matrix, given by

$$\begin{aligned} \mathrm{Var}(\widehat{\beta}) &= \mathrm{Var}\{(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y\} && (1.6)\\ &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathrm{Var}(Y)X(X^{\mathrm{T}}X)^{-1} && (1.7)\\ &= (X^{\mathrm{T}}X)^{-1}\sigma^2\,, && (1.8) \end{aligned}$$

where $Y \sim N(X\beta, I_n\sigma^2)$, where I_n is an $n \times n$ identity matrix.

[5] Check the Matrix Cookbook for matrix calculus, amongst much else.

Hence, in our example,

$$\begin{aligned}\mathrm{Var}(\hat{\beta}) &= \frac{1}{n\sum x_j^2 - (\sum x_j)^2}\begin{bmatrix} \sum x_j^2 & -\sum x_j \\ -\sum x_j & n \end{bmatrix}\sigma^2 \\ &= \begin{bmatrix} \mathrm{Var}(\hat{\beta}_0) & \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \mathrm{Var}(\hat{\beta}_1) \end{bmatrix}.\end{aligned}$$

For estimating the difference between treatments, we are interested in

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_1) &= \frac{n}{n\sum x_j^2 - (\sum x_j)^2}\sigma^2 \\ &= \frac{n}{n^2 - (\sum x_j)^2}\sigma^2\,,\end{aligned}$$

as $x_j = \pm 1$, therefore $x_j^2 = 1$ for all $j = 1, \ldots, n$, and hence $\sum x_j^2 = n$.

To achieve the most precise estimator, we need to minimise $\mathrm{Var}(\hat{\beta}_1)$ or equivalently minimise $|\sum x_j|$. This goal can achieve this through the choice of $x_1, \ldots, x_N$:

- as each x_j can only take one of two values, -1 or +1, this is equivalent to choosing the numbers of subjects assigned to treatment 1 and treatment 2;
- call these n_1 and n_2 respectively, with $n_1 + n_2 = N$

It is obvious that $\sum x_j = 0$ if and only if $n_1 = n_2$. Therefore, assuming N is even, the **optimal design** has

- $n_1 = \frac{n}{2}$ subjects assigned to treatment 1 and
- $n_2 = \frac{n}{2}$ subjects assigned to treatment 2.

For N odd, we choose $n_1 = \frac{n+1}{2}$, $n_2 = \frac{n-1}{2}$, or vice versa.

Definition 1.3. We can assess different designs using their **efficiency**:

$$\mathrm{Eff} = \frac{\mathrm{Var}(\hat{\beta}_1 \,|\, d^*)}{\mathrm{Var}(\hat{\beta}_1 \,|\, d_1)} \tag{1.9}$$

where d_1 is a design we want to assess and d^* is the optimal design with smallest variance. Note that $0 \leq \mathrm{Eff} \leq 1$.

In Figure 1.2 below, we plot this efficiency for Example 1.1, using different choices of n_1. The total number of runs is fixed at $n = 100$, and the function `eff` calculates the efficiency from Definition 1.3 for a design with n_1 subjects assigned to treatment 1. Clearly, efficiency of 1 is achieved when $n_1 = n_2$ (equal allocation of treatments 1 and 2). If $n_1 = 0$ or $n_1 = 1$, the efficiency is zero; we cannot estimate the difference between two treatments if we only allocate subjects to one of them.

```
n <- 100
eff <- function(n1) 1 - ((2 * n1 - n) / n)^2
curve(eff, from = 0, to = n, ylab = "Eff", xlab = expression(n[1]))
```

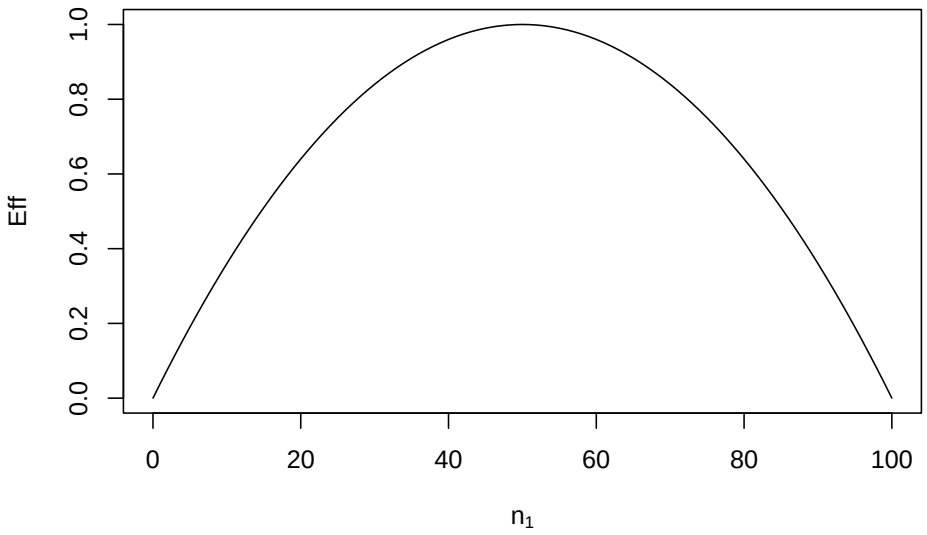

Figure 1.2: Efficiencies for designs for Example 1.1 with different numbers, n_1, of subjects assigned to treatment 1 when the total number of subjects is $n = 100$.

1.2 Aims of experimentation and some examples

Some reasons experiments are performed:

1. Treatment comparison (Chapters 2 and 3)

- compare several treatments (and choose the best)
- e.g. clinical trial, agricultural field trial

2. Factor screening (Chapters 4, 5 and 6)

- many complex systems may involve a large number of (discrete) factors (controllable features)
- which of these factors have a substantive impact?
- (relatively) small experiments
- e.g. industrial experiments on manufacturing processes

3. Response surface exploration (Chapter 7)

- detailed description of relationship between important (continuous) variables and response

- typically second order polynomial regression models
- larger experiments, often built up sequentially
- e.g. alcohol yields in a pharmaceutical experiments

4. Optimisation (Chapter 7)

- finding settings of variables that lead to maximum or minimum response
- typically use response surface methods and sequential "hill climbing'' strategy

1.3 Some definitions

Definition 1.4. The **response** Y is the outcome measured in an experiment; e.g. yield from a chemical process. The response from the n observations are denoted $Y_1, \ldots, Y_n$.
Definition 1.5. Factors (discrete) or **variables** (continuous) are features which can be set or controlled in an experiment; m denotes the number of factors or variables under investigation. For discrete factors, we call the possible settings of the factor its **levels**. We denote by x_{ij} the value taken by factor or variable i in the jth run of the experiment ($i = 1, \ldots, m$; $j = 1, \ldots, n$).
Definition 1.6. The **treatments** or **support points** are the *distinct* combinations of factor or variable values in the experiment.
Definition 1.7. An experimental **unit** is the basic element (material, animal, person, time unit, …) to which a treatment can be applied to produce a response.

In Example 1.1 (comparing two treatments):

- Response Y: Measured outcome, e.g. protein level or pain score in clinical trial, yield in an agricultural field trial.
- Factor x: "treatment'' applied
- Levels

  treatment 1 $x = -1$
  treatment 2 $x = +1$

- Design point: factor level applied to jth subject; $x_j = \pm 1$
- Treatment or support point: Two treatments or support points
- Experimental unit: Subject (person, animal, plot of land, …).

1.4 Principles of experimentation

Three fundamental principles that need to be considered when designing an experiment are:

- replication
- randomisation
- stratification (blocking)

1.4.1 Replication

Each treatment is applied to a number of experimental units, with the jth treatment replicated r_j times. This enables the estimation of the variances of treatment effect estimators; increasing the number of replications, or replicates, decreases the variance of estimators of treatment effects. (Note: proper replication involves independent application of the treatment to different experimental units, not just taking several measurements from the same unit).

1.4.2 Randomisation

Randomisation should be applied to the allocation of treatments to units. Randomisation protects against **bias**; the effect of variables that are unknown and potentially uncontrolled or subjectivity in applying treatments. It also provides a formal basis for inference and statistical testing.

For example, in a clinical trial to compare a new drug and a control random allocation protects against

- "unmeasured and uncontrollable" features (e.g. age, sex, health)
- bias resulting from the clinician giving new drug to patients who are sicker.

Clinical trials are usually also *double-blinded*, i.e. neither the healthcare professional nor the patient knows which treatment the patient is receiving.

1.4.3 Stratification (or blocking)

We would like to use a wide variety of experimental units (e.g. people or plots of land) to ensure **coverage** of our results, i.e. validity of our conclusions across the population of interest. However, if the sample of units from the population is too heterogenous, then this will induce too much random variability, i.e. increase σ^2 in $\varepsilon_j \sim N(0, \sigma^2)$, and hence increase the variance of our parameter estimators.

We can reduce this extraneous variation by splitting our units into homogenous sets, or **blocks**, and including a blocking term in the model. The simplest blocked experiment is a **randomised complete block design**, where each block contains enough units for all treatments to be applied. Comparisons can then be made *within* each block.

Basic principle: block what you can, randomise what you cannot.

Later we will look at blocking in more detail, and the principle of **incomplete blocks**.

1.5 Revision on the linear model

Recall: $Y = X\beta + \varepsilon$, with $\varepsilon \sim N(0, I_n\sigma^2)$. Let the jth row of X be denoted x_j^{T}, which holds the values of the predictors, or explanatory variables, for the jth observation. Then

$$Y_j = x_j^{\mathrm{T}}\beta + \varepsilon_j\,, \quad j = 1, \ldots, n\,.$$

For example, quite commonly, for continuous variables

$$x_j = (1, x_{1j}, x_{2j}, \ldots, x_{mj})^{\mathrm{T}}\,,$$

and so

$$x_j^{\mathrm{T}}\beta = \beta_0 + \beta_1 x_{1j} + \cdots + \beta_m x_{mj}\,.$$

The laest squares estimators are given by

$$\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y\,,$$

with

$$\mathrm{Var}(\hat{\beta}) = (X^{\mathrm{T}}X)^{-1}\sigma^2\,.$$

1.5.1 Variance of a Prediction/Fitted Value

A prediction of the mean response at point x_0 (which may or may not be in the design) is

$$\hat{Y}_0 = x_0^{\mathrm{T}}\hat{\beta}\,,$$

with

$$\begin{aligned}\mathrm{Var}(\hat{Y}_0) &= \mathrm{Var}\left(x_0^{\mathrm{T}}\hat{\beta}\right)\\ &= x_0^{\mathrm{T}}\mathrm{Var}(\hat{\beta})x_0\\ &= x_0^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_0\sigma^2\,.\end{aligned}$$

For a linear model, this variance depends only on the assumed regression model and the design (through X), the point at which prediction is to be made (x_0) and the value of σ^2; it does not depend on data Y or parameters β.

Similarly, we can find the variance-covariance matrix of the fitted values:

$$\mathrm{Var}(\hat{Y}) = \mathrm{Var}(X\hat{\beta}) = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\sigma^2\,.$$

1.5.2 Analysis of Variance and $\mathbf{R}^2$ as Model Comparison

To assess the goodness-of-fit of a model, we can use the residual sum of squares

$$\begin{aligned}\mathrm{RSS} &= (Y - X\widehat{\beta})^{\mathrm{T}}(Y - X\widehat{\beta}) \\ &= \sum_{j=1}^{n} \left\{Y_j - x_j^{\mathrm{T}}\widehat{\beta}\right\}^2 \\ &= \sum_{j=1}^{n} r_j^2 \,,\end{aligned}$$

where

$$r_j = Y_j - x_j^{\mathrm{T}}\widehat{\beta}\,.$$

Often, a comparison is made to the null model

$$Y_j = \beta_0 + \varepsilon_j \,,$$

i.e. $Y_i \sim N(\beta_0, \sigma^2)$. The residual sum of squares for the null model is given by

$$\mathrm{RSS(null)} = Y^{\mathrm{T}}Y - m\bar{Y}^2 \,,$$

as

$$\widehat{\beta}_0 = \bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j \,.$$

How do we compare these models?

1. Ratio of residual sum of squares:

$$\begin{aligned}R^2 &= 1 - \frac{\mathrm{RSS}}{\mathrm{RSS(null)}} \\ &= 1 - \frac{(Y - X\widehat{\beta})^{\mathrm{T}}(Y - X\widehat{\beta})}{Y^{\mathrm{T}}Y - n\bar{Y}^2} \,.\end{aligned}$$

The quantity $0 \leq R^2 \leq 1$ is sometimes called the **coefficient of multiple determination**:

- high R^2 implies that the model describes much of the variation in the data;

- **but** note that R^2 will always increase as p (the number of explanatory variables) increases, with $R^2 = 1$ when $p = n$;
- some software packages will report the adjusted R^2.

$$\begin{aligned} R_a^2 &= 1 - \frac{\text{RSS}/(n-p)}{\text{RSS(null)}/(n-1)} \\ &= 1 - \frac{(Y - X\hat{\beta})^{\mathrm{T}}(Y - X\hat{\beta})/(n-p)}{(Y^{\mathrm{T}}Y - n\bar{Y}^2)/(n-1)}; \end{aligned}$$

- R_a^2 does not necessarily increase with p (as we divide by degrees of freedom to adjust for complexity of the model).

2. Analysis of variance (ANOVA): An ANOVA table is compact way of presenting the results of (sequential) comparisons of nested models. You should be familiar with an ANOVA table of the following form.

Table 1.1: A standard ANOVA table.

Source	Degress of Freedom	(Sequential) Sum of Squares	Mean Square
Regression	$p-1$	By subtraction; see (1.12)	Reg SS$/(p-1)$
Residual	$N-p$	$(Y - X\hat{\beta})^{\mathrm{T}}(Y - X\hat{\beta})$[6]	RSS$/(N-p)$
Total	$N-1$	$Y^{\mathrm{T}}Y - N\bar{Y}^2$[7]	

In row 1 of Table 1.1 above,

$$\text{Regression SS} = \text{Total SS} - \text{RSS} = Y^{\mathrm{T}}Y - n\bar{Y}^2 - (Y - X\hat{\beta})^{\mathrm{T}}(Y - X\hat{\beta}) \tag{1.10}$$

$$= -n\bar{Y}^2 - \hat{\beta}^{\mathrm{T}}(X^{\mathrm{T}}X)\hat{\beta} + 2\hat{\beta}^{\mathrm{T}}X^{\mathrm{T}}Y \tag{1.11}$$

$$= \hat{\beta}^{\mathrm{T}}(X^{\mathrm{T}}X)\hat{\beta} - n\bar{Y}^2\,, \tag{1.12}$$

with the last line following from

$$\begin{aligned} \hat{\beta}^{\mathrm{T}}X^{\mathrm{T}}Y &= \hat{\beta}^{\mathrm{T}}(X^{\mathrm{T}}X)(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y \\ &= \hat{\beta}^{\mathrm{T}}(X^{\mathrm{T}}X)\hat{\beta} \end{aligned}$$

[6] Residual sum of squares for the full regression model.
[7] Residual sum of squares for the null model.

This idea can be generalised to the comparison of a *sequence* of nested models - see Problem Sheet 1.

Hypothesis testing is performed using the mean square:

$$\frac{\text{Regression SS}}{p-1} = \frac{\hat{\beta}^{\mathrm{T}}(X^{\mathrm{T}}X)\hat{\beta} - n\bar{Y}^2}{p-1}\,.$$

Under $\mathrm{H}_0 : \beta_1 = \cdots = \beta_{p-1} = 0$

$$\begin{aligned}\frac{\text{Regression SS}/(p-1)}{\text{RSS}/(N-p)} &= \frac{(\hat{\beta}^{\mathrm{T}}(X^{\mathrm{T}}X)\hat{\beta} - n\bar{Y}^2)/(p-1)}{(Y - X\hat{\beta})^{\mathrm{T}}(Y - X\hat{\beta})/(n-p)} \\ &\sim F_{p-1,n-p}\,,\end{aligned}$$

an F distribution with $p-1$ and $n-p$ degrees of freedom; defined via the ratio of two independent χ^2 distributions.

Also,

$$\frac{\text{RSS}}{n-p} = \frac{(Y - X\hat{\beta})^{\mathrm{T}}(Y - X\hat{\beta})}{n-p} = \hat{\sigma}^2$$

is an unbiased estimator for σ^2, and

$$\frac{(n-p)}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-p}\,.$$

This is a Chi-squared distribution with $N-p$ degrees of freedom (see MATH2010 notes).

Chapter 2

Simple comparative experiments

Bibliography

Luca, M. and Bazerman, M. H. (2020). *The Power of Experiments: Decision Making in a Data-Driven World*. MIT Press, Cambridge, MA.