Optimal design of experiments

Session 1: Introduction

Peter Goos

Universiteit Antwerpen

Purpose of experimentation

- quantify relationship between some response(s) and one or more explanatory / experimental variables
- involves changing the system under study and observing the effect changes have on the system (↔ observational study or survey)
- advantages:
  - draw causal inferences rather than note patterns
  - informative events can be made to happen
  - yields the data that are needed

Purpose of this course

- there are huge libraries with lists of experimental designs
- practical problems rarely allow one of these to be used without any change
- people often change their problem to fit the experimental design
- this course is about creating the best possible design for a given problem

Example from industry

- response y = voltage output
- depends on $\begin{cases} \text{blade speed} & x_1 \\ \text{extension} & x_2 \end{cases}$
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon$
- linear model in β-parameters

Example from industry

- $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$
  $+\beta_{12} x_{1i} x_{2i} + \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 + \epsilon_i$
- data

run	voltage	speed	extension
1	1.23	5300	0.000
2	3.13	8300	0.000
3	1.22	5300	0.012
⋮	⋮	⋮	⋮
11	1.59	6800	0.006

treatment

Examples from medicine and psychology

- medicine
  - response y = corneal hydration
  - depends on CO_2 level x
  - $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$
  - two treatments per subject
- psychology
  - response y = number of mistakes on a test
  - depends on number of hours test person is awake
  - $y = \beta_0 + \beta_1 x + \epsilon$

Choice experiment

- response y = race bicycle that is bought
- depends on
  $$\begin{cases} \text{type of frame} & x_1, x_2 \\ \text{brand of gears and brakes} & x_3 \\ \text{type of wheels} & x_4 \end{cases}$$
- utility $U = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$
- multinomial logit model

$$P(\text{option 1 is chosen}) = \frac{e^{\beta_0 + \beta_1 x_{11} + \ldots}}{\sum_i e^{\beta_0 + \beta_1 x_{1i} + \ldots}}$$

- nonlinear in the β-parameters

Marketing experiment

Which of the two race bicycles would you prefer if the options only differ with respect to the attributes shown?	
	
Carbon frame	Aluminium frame
Classic frame	Sloping frame
Mavic Ksyrium SL wheels	Shimano WH-7701 wheels
Campagnolo Record groupset	Shimano Dura-Ace groupset

Rating-based conjoint experiment

- response y = willingness-to-pay for a bicycle
- depends on

$$\begin{cases} \text{type of frame} & x_1, x_2 \\ \text{brand of gears and brakes} & x_3 \\ \text{type of wheels} & x_4 \end{cases}$$

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$

Models and variables

- variables = factors (engineers call these parameters)
- quantitative variables vs. qualitative (categorical) ones
- linear models vs. *non-linear* models *(non-linear in the unknown model parameters)*
- most examples involve quantitative variables but methodology can easily handle qualitative variables too
- first: linear models!

Example from industry

- $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$
  $+ \beta_{12} x_{1i} x_{2i} + \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 + \epsilon_i$
- data

run	voltage	speed	extension
1	1.23	5300	0.000
2	3.13	8300	0.000
3	1.22	5300	0.012
⋮	⋮	⋮	⋮
11	1.59	6800	0.006

treatment

Variable scaling

run	y voltage	x_1 speed	x_2 extension
1	1.23	-1	-1
2	3.13	$+1$	-1
3	1.22	-1	$+1$
⋮	⋮	⋮	⋮
11	1.59	0	0

$\downarrow$

$x = \frac{u - u_0}{\Delta}$

where u = original value, u_0 = midpoint between min and max, Δ = half the interval

Assumption of independence

- $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$
  $+ \beta_{12} x_{1i} x_{2i} + \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 + \epsilon_i$
- order of experimental runs is randomized
- make sure responses are independent
  (e.g. reset factor levels for every run)
- all ϵ_i independent
- $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$
  $\Rightarrow$ ordinary least squares (OLS) is best linear unbiased estimator

OLS estimator

- $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}$

  where

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 & \hat{\beta}_1 & \hat{\beta}_2 & \hat{\beta}_{12} & \hat{\beta}_{11} & \hat{\beta}_{22} \end{bmatrix}^T$$

$$\mathbf{X} = \begin{array}{c} \begin{bmatrix} 1 & -1 & -1 & +1 & +1 & +1 \\ 1 & +1 & -1 & -1 & +1 & +1 \\ & & & \vdots & & \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\ \begin{array}{cccccc} \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\ \text{int.} & x_1 & x_2 & x_1 x_2 & x_1^2 & x_2^2 \end{array} \end{array}$$

$$\mathbf{y} = \begin{bmatrix} 1.23 & 3.13 & 1.22 & \dots & 1.59 \end{bmatrix}^T$$

Estimated model

- estimate of factor effects

$$\mathbf{b} = \begin{bmatrix} 1.67 & 0.65 & -0.29 & -0.30 & 0.22 & 0.02 \end{bmatrix}^T$$

- estimated model

$$\begin{aligned}\hat{y}_i &= 1.67 + 0.65x_1 + (-0.29)x_2 + (-0.30)x_1x_2 \\ &\qquad + 0.22x_1^2 + 0.02x_2^2 \\ &= 1.67 + 0.65x_1 - 0.29x_2 - 0.30x_1x_2 \\ &\qquad + 0.22x_1^2 + 0.02x_2^2 \\ &= \mathbf{f}^T(\mathbf{x}_i)\,\mathbf{b}\end{aligned}$$

where $\mathbf{f}^T(\mathbf{x}_i) = \begin{bmatrix} 1 & x_{1i} & x_{2i} & x_{1i}x_{2i} & x_{1i}^2 & x_{2i}^2 \end{bmatrix}$

Inference

- variance-covariance matrix of $\hat{\beta}$

$$\text{var}(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

- estimate σ^2 using mean squared error

$$\text{MSE} = \frac{\mathbf{r}^T\mathbf{r}}{n-p} \begin{array}{l} \rightarrow \text{sum of squared residuals} \\ \rightarrow \text{residual degrees of freedom} \end{array}$$

where

$$\begin{aligned}\mathbf{r} &= \mathbf{y} - \mathbf{Xb} \\ n &= \#\ \text{observations} \\ p &= \#\ \text{model parameters}\end{aligned}$$

Variance-covariance matrix

$$\text{var}(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 0.26 & 0 & 0 & 0 & -0.16 & -0.16 \\ 0 & 0.17 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.17 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.25 & 0 & 0 \\ -0.16 & 0 & 0 & 0 & 0.39 & -0.11 \\ -0.16 & 0 & 0 & 0 & -0.11 & 0.39 \end{bmatrix}$$

Information matrix

$$\frac{1}{\sigma^2}\left(\mathbf{X}^T\mathbf{X}\right) = \begin{bmatrix} 11 & 0 & 0 & 0 & 6 & 6 \\ 0 & 6 & 0 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 \\ 6 & 0 & 0 & 0 & 6 & 4 \\ 6 & 0 & 0 & 0 & 4 & 6 \end{bmatrix}$$

(diagonal contains “effective sample sizes”)

- point prediction

$$\begin{aligned}\hat{y}_i &= 1.67 + 0.65x_1 - 0.29x_2 - 0.30x_1x_2 \\ &\qquad + 0.22x_1^2 + 0.02x_2^2 \\ &= \mathbf{f}^T(\mathbf{x}_i)\,\mathbf{b}\end{aligned}$$

- prediction variance

$$\mathrm{var}\left(\hat{y}_i\right) = \sigma^2 \mathbf{f}^T(\mathbf{x}_i)\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{f}(\mathbf{x}_i)$$