Biostat 279: Lecture 3

Outline:

- Research steps in carrying out a study

- Design defined and issues

- Review of optimality ideas in last lecture
  with a focus on simple linear model

Introduction to optimal design ideas

Recommended Readings:

Book by Atkinson, A.C. & Donev, A.N. (1992) Optimum experimental Designs,Oxford: Clarendon Press

Paper by Wong, W. K. and Lachenbruch, P. A.(1996). Designing Studies for Dose Response. Statistics in Medicine, Vol. 15, 343-360. (Tutorial in Biostatistics)

Research steps

- Formulation of problem
- Research design/experiment
- Choice of statistical model
- Analysis of the data
- Conclusions

dependency

Design Considerations

- Independent variables (factors) and their levels or range of values
- Dependent variable and choice of measurement
- Assignment of units (subjects) to combinations of levels
- Sample size

Observational studies in a picture

	Ex-posed	Unex-posed	Total
Case			
Non-case			
Total			

Case-control Study: fixed margins

X-sectional Study: fixed grand total

Cohort Study: fixed margins

What is a design?

Chararteristics of independent variables:

- Levels may be quantitative or qualitative
- Levels may be fixed or random
- Variables may be crossed or nested
- Weights (replications) for each combination of levels may be equal or not

What is a good design?

A good design leads to:

- Maximum statistical inference precision
- Minimum cost possible (budget, time, labor)
- Produce results that people have faith in

A very brief history

- Fisher (1935): Agricultural experiments, latin squares, blocked designs

- Cox (1958): Planning of experiments

- Kiefer (1959), Box and Draper (1971)

- Federov (1972), Silvey (1980), Atkinson and Donev (1992)

Some requirements for proper experimentation Box & Draper (1971):

- Generate sufficient information across the region of interest
- Require a minimum number of observations/runs
- Avoid large differences in number of levels
- Obtain good estimates of effects, error variance and check assumptions

Some requirements for proper experimentation Box & Draper (1971):

- Allow designs to be built sequentially
- Allow for Blocking
- Ensure good detection procedures for lack-of-Fit
- Keep things simple

2 Two scenarios for an optimal design:

1. Given costs of experimentation and certain power, compute optimal sample size.

2. Alternatively, sample size is predetermined and we decide on an optimal allocation scheme.

Advantages of optimal designs

- Optimal designs are based on more or less some optimality criteria – which may be subjective or difficult to choose
- The optimal design methodology provides possibility to build a design sequentially
- General guidelines for allocating resources and may be used as a gold standard

Disadvantages of optimal designs

- The statistical model has to be specified in advance
- Optimal designs generally depend on the optimality criterion and all aspects of the model
- Optimal designs are sometimes not easy to find and implement

Psychotherapeutic experiment

y= Excessive fear or phobic reaction(continuous)

x: x=0: Standard therapy (A): $y_i = \beta_0 + \varepsilon_i$

x =1: Innovative therapy (B): $y_i = \beta_0 + \beta_1 + \varepsilon_i$

Two random samples of patients, n_A= 3 x n_B

Psychotherapeutic experiment

How many patients *N*? ($n_A = \frac{3}{4} N$ $n_B = \frac{1}{4} N$)

$$t = \frac{M_A - M_B - (\mu_A - \mu_B)}{S_p \sqrt{\frac{n_A n_B}{n_A + n_B}}}, \quad \alpha = 0.05, \quad \delta = 0.8$$

$$(1 - \beta) = 0.85 \ \ ?!$$

Psychotherapeutic experiment

$$\alpha = 0.05, \quad \delta = (M_A - M_B) / \sigma = 0.8$$

n_A	n_B	Power	n_A	n_B	Power
15	5	0.27	10	10	0.39
30	10	0.56	20	20	0.69
45	15	0.73	30	30	0.86
60	20	0.86	40	40	0.94

Psychotherapeutic experiment

$c_B = 3\ c_A$

n_A	n_B	Costs	Power	n_A	n_B	Costs	Power
15	5	$30c_A$	0.27	10	10	$40c_A$	0.39
30	10	$60c_A$	0.56	20	20	$80c_A$	0.69
45	15	$90c_A$	0.73	30	30	$120c_A$	0.86
60	20	$120c_A$	0.86	40	40	$160c_A$	0.94

Dose response experiment

y= Tumor shrinkage (quantitative)
x= Radiation dosages (quantitative)

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

y_i = response of subject i

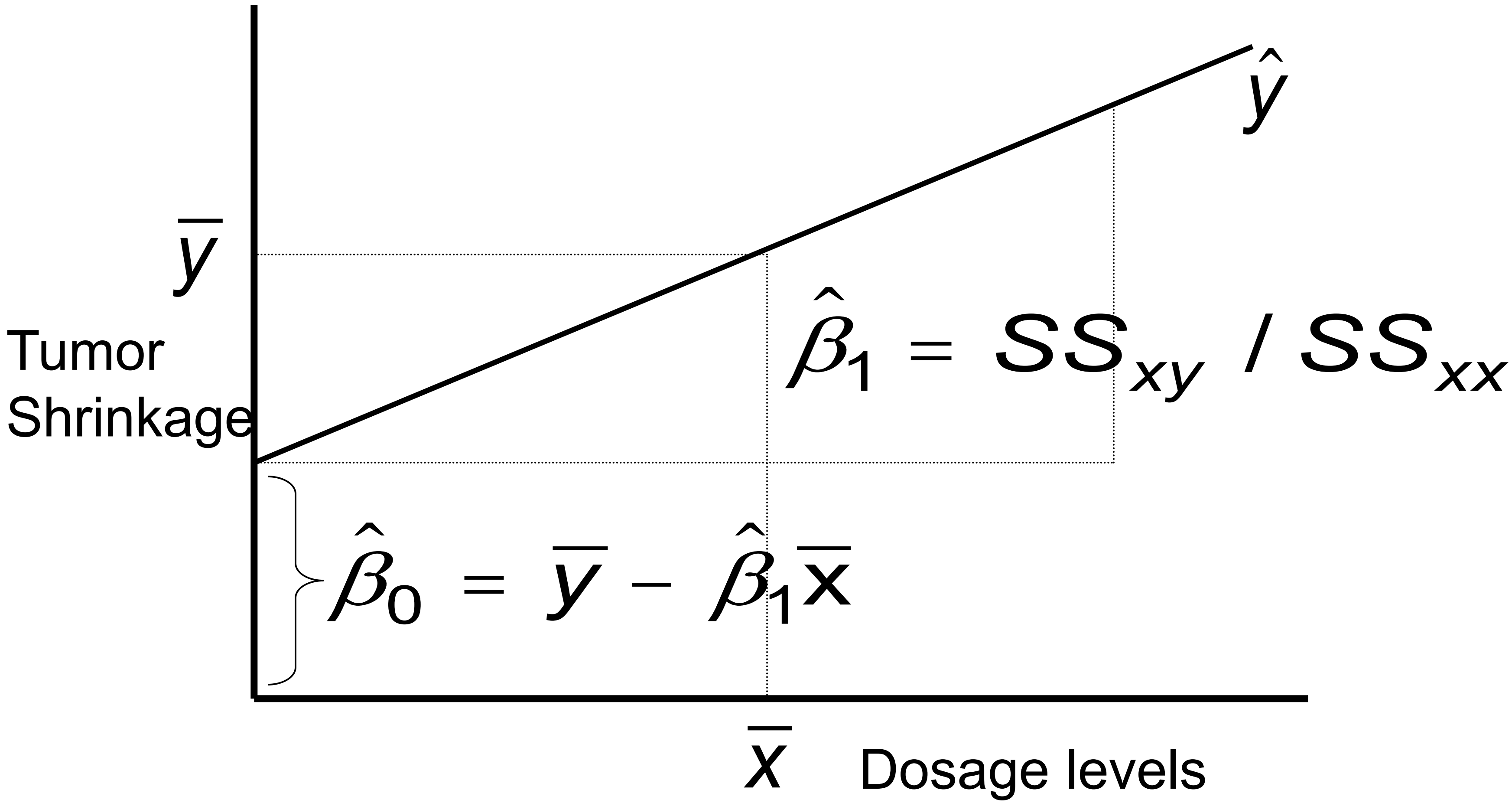
β_0 = Intercept

β_1 = slope

x_i = Independent variable

ε_i = errors: $N(o, \sigma_\varepsilon^2)$

Linear regression



Linear regression

OLS

$\hat{\beta}_0$: $$\mathrm{Var}(\hat{\beta}_0) = \sigma_\varepsilon^2\left(\frac{1}{N} + \frac{\bar{\mathrm{x}}^2}{\mathrm{SS}_{\mathrm{xx}}}\right)$$

Covariance

$\hat{\beta}_1$: $$Var(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{SS_{xx}}$$

Linear regression

$$OLS \;\; \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$$

$$Cov(\hat{\beta}) = \frac{\sigma_\varepsilon^2}{NSS_{xx}} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & N \end{bmatrix}$$

Linear regression

$$OLS \;\; \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$$

$$Cov(\hat{\beta}) = \frac{\sigma_\varepsilon^2}{NSS_{xx}} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & N \end{bmatrix}$$

Linear regression

$$OLS \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$$

$$Cov(\hat{\beta}) = \frac{\sigma_\varepsilon^2}{NSS_{xx}} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & N \end{bmatrix}$$

Linear regression

Confidence interval β_1

$$\hat{\beta}_1 - t_{\alpha/2, N-2}\sqrt{\frac{MS_e}{SS_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2, N-2}\sqrt{\frac{MS_e}{SS_{xx}}}$$

Linear regression

Confidence interval β_0

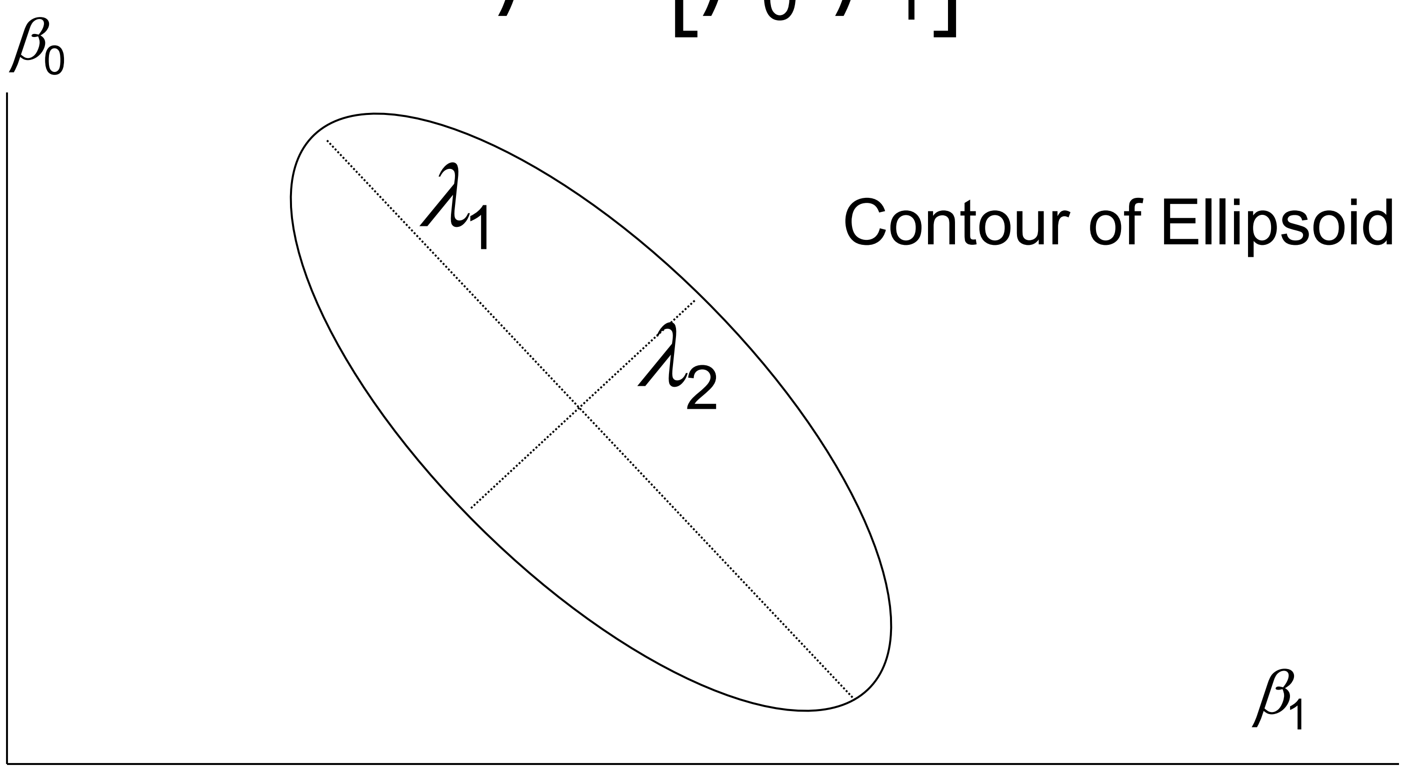
$$\hat{\beta}_0 - t_{\alpha/2,N-2}\sqrt{MS_e\left(\frac{1}{N}+\frac{\overline{x}^2}{SS_{xx}}\right)} \le \beta_0$$

$$\le \hat{\beta}_0 + t_{\alpha/2,N-2}\sqrt{MS_e\left(\frac{1}{N}+\frac{\overline{x}^2}{SS_{xx}}\right)}$$

Linear regression

Simultaneous confidence interval

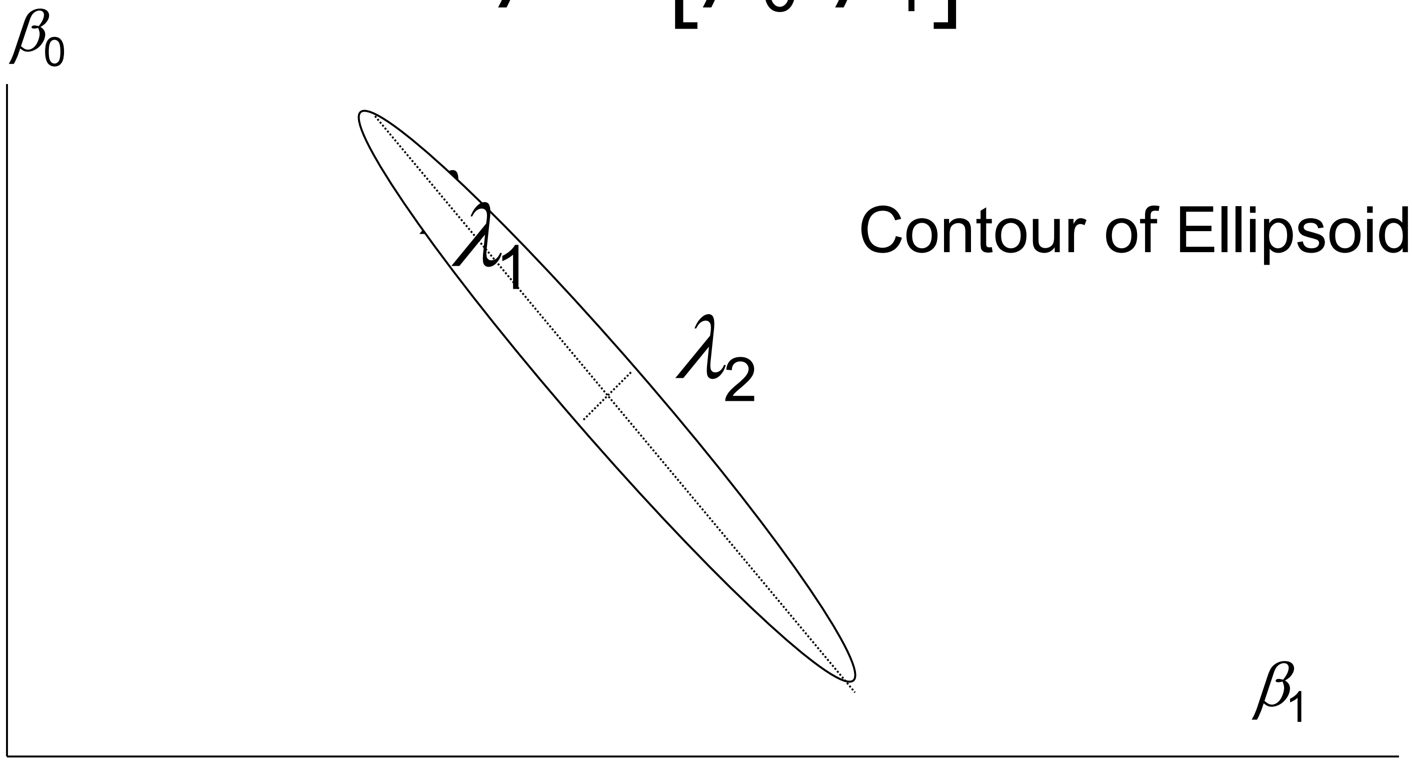
$$\beta = \left[\beta_0\ \beta_1\right]$$



Linear regression

Simultaneous confidence interval

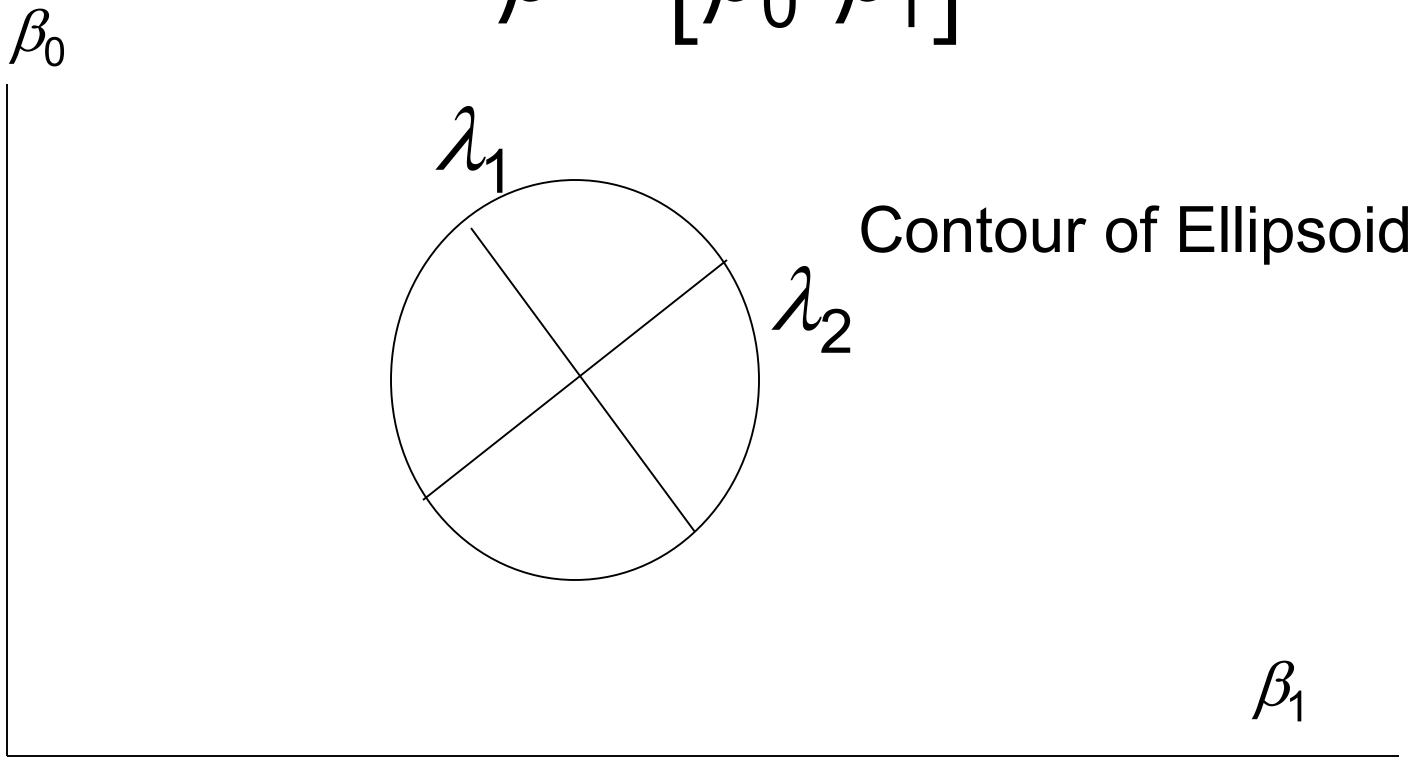
$$\beta = [\beta_0 \ \beta_1]$$



Linear regression

Simultaneous confidence interval

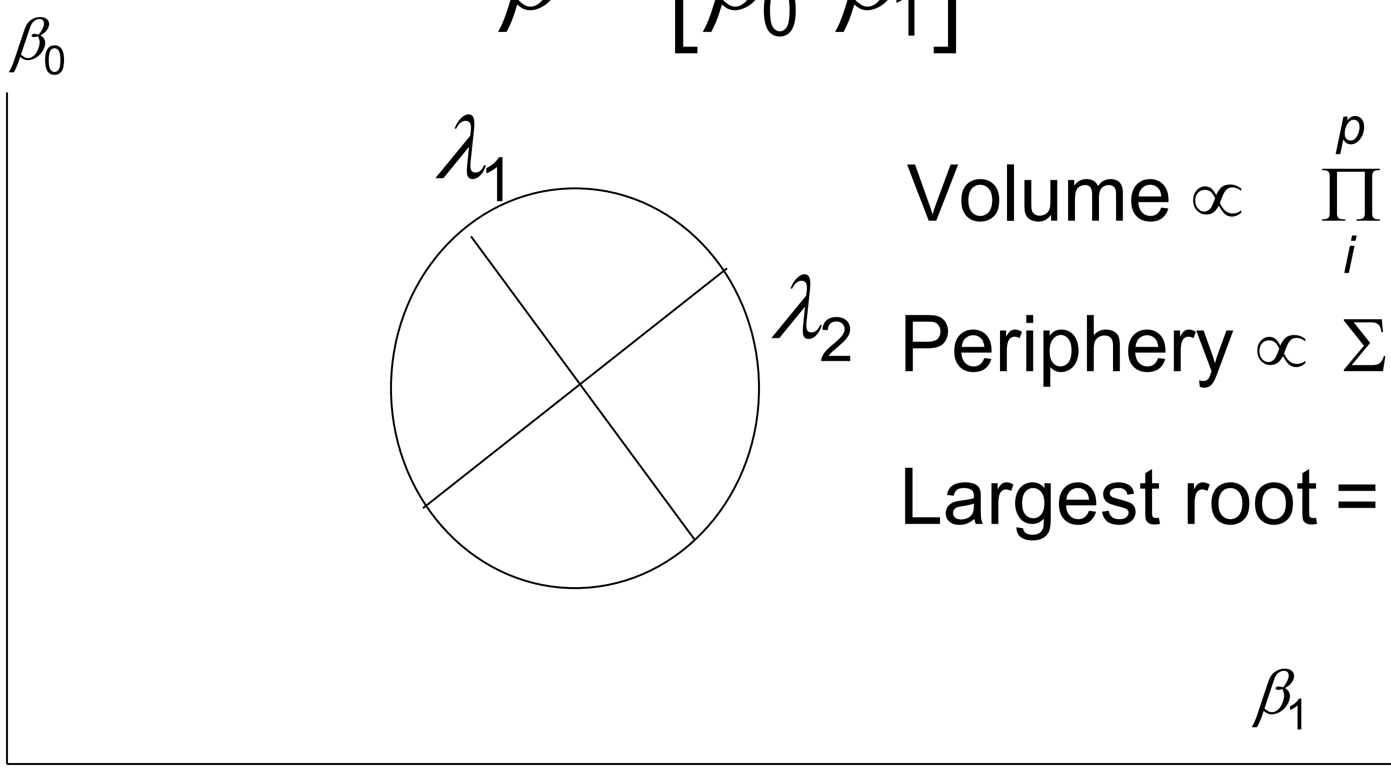
$$\beta = \left[\beta_0 \ \beta_1\right]$$



Linear regression

Characteristics of ellipsoid

$$\beta = \left[\beta_0 \; \beta_1\right]$$



Volume $\propto \prod_{i}^{p} \lambda_i$

Periphery $\propto \sum_{i}^{p} \lambda_i$

Largest root $= \max_{i}(\lambda_i)$

Linear regression

Simultaneous confidence interval

$$\beta = \left[\beta_0 \ \beta_1\right]$$

$$\mathrm{Prob}\left(\frac{\mathrm{Ellipse}}{2MS_e} \le F_{\alpha,2,N-2} \right) = 1 - \alpha$$

$$\begin{aligned}\mathrm{Ellipse} = N(\hat{\beta}_0 - \beta_0)^2 &+ \\ 2\Sigma x_i(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1) &+ \\ \Sigma x_i^2(\hat{\beta}_1 - \beta_1)^2 &\end{aligned}$$

Linear regression

Matrix formulation

$$y = X\beta + e$$

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ 1. & x_k \\ . & . \\ 1 & x_N \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_N \end{bmatrix}$$

Linear regression

Summary *OLS*:

$$\hat{\beta} = (X'X)^{-1}X'y$$

$$\mathrm{Cov}(\hat{\beta}) = \sigma_{\varepsilon}^{2}(X'X)^{-1}$$

$$\hat{\sigma}_{\varepsilon}^{2} = \left[y'y - \hat{\beta}X'y\right] / (N - p)$$

Linear regression

Simultaneous confidence interval

$$\beta = \left[\beta_0\ \beta_1\ \beta_2\ \ldots \beta_{p-1}\right]$$

$$\text{Prob}\left(\frac{(N - p)(\hat{\beta} - \beta)'(\text{X}'\text{X})(\hat{\beta} - \beta)}{p(y'y - \hat{\beta}'\text{X}'\text{y})} \le F_{\alpha,p,N-p} \right) = 1 - \alpha$$

Exact Design measure

$$y = X\beta + e$$

Discrete design measure

$$\xi = \begin{Bmatrix} x_1 & x_2 & x_3 & . & . & . & x_k \\ n_1 & n_2 & n_3 & . & . & . & n_k \end{Bmatrix}, \; x_j \in \chi, \; \sum_j^k n_j = N$$

$$\xi = \begin{Bmatrix} 10 & 20 & 30 & . & . & . & 60 \\ 5 & 5 & 5 & . & . & . & 5 \end{Bmatrix}, \; x_j \in \{10,60\}, \; 5k = N$$

Approximate Design measure

$$y = X\beta + e$$

Continuous (approximate) design measure

$$\xi = \begin{Bmatrix} x_1 & x_2 & x_3 & . & . & . & x_k \\ w_1 & w_2 & w_3 & . & . & . & w_k \end{Bmatrix}, \; x_j \in \chi, \;\; 0 \le \mathrm{w}_j \le 1$$

$$\int_{\chi} \xi(dx) = 1.$$

The approximate relationship is $Nw_i = n_i$, $i = 1,2,\ldots, k$.

Optimality criteria

- Fisher information matrix: $M = (X'X)$
- Asymptotic variance of estimators

$$\mathrm{Cov}(\hat{\beta}) = M^{-1} = (X'X)^{-1}$$

– $\Psi\{M(\xi^*)\} = \max_{\xi \in \chi} \Psi\{M(\xi)\}$

– $\Psi\{M(\xi^*)^{-1}\} = \min_{\xi \in \chi} \Psi\{M(\xi)^{-1}\}$

Optimality criteria

$$\text{D-optimality:} \quad \min_{\xi \in \chi}\{Det[X'X)^{-1}]$$

Advantages:

- D-optimality criterion is proportional to volume of confidence ellipsoid, thus having natural interpretation $\propto \prod_{i}^{p} \lambda_i$
- D-optimal design generally perform well compared to other criteria (e.g. Donev and Atkinson, 1988)

Optimality criteria

$$\text{D - optimality:} \quad \min_{\xi \in \chi}\{Det[X'X)^{-1}]$$

Advantages:

- D-optimal designs are invariant under linear transformation of the design matrix

$$Z = XT \qquad (T \text{ is nonsingular})$$

$$|Z'Z| = |(XT)'(XT)| = |T'(X'X)T| = |T'T|\ |X'X|$$

Estimating a Subset of the Model Parameters

Suppose β is partitioned into β_0 and β_1.

$$D_s\text{ - optimality:}\quad \mathrm{Cov}(\hat{\beta}) = \begin{bmatrix} \mathrm{Var}(\hat{\beta}_0) & \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \mathrm{Var}(\hat{\beta}_1) \end{bmatrix}$$

$$(X'X) = \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_1'X_2 & X_2'X_2 \end{bmatrix} \qquad (X'X)^{-1} = \begin{bmatrix} \tilde{X}_1'\tilde{X}_1 & \tilde{X}_1'\tilde{X}_2 \\ \tilde{X}_1'\tilde{X}_2 & \tilde{X}_2'\tilde{X}_2 \end{bmatrix}$$

$$\min_{\xi \in \chi}\{Det[\tilde{X}_1'\tilde{X}_1]\}$$

where

$$\tilde{X}_1'\tilde{X}_1 = [X_1'X_1 - X_1'X_2(X_2'X_2)^{-1}X_2'X_1]^{-1}$$

Optimality criteria

D - optimality with or without intercept β_0 is equivalent:

$$Det[X_1' X_1 - X_1' X_2 (X_2' X_2)^{-1} X_2' X_1] = \frac{Det[X' X]}{Det[X_2' X_2]}$$

where $Det(\mathrm{X}_2' X_2) = Det(1_N' 1_N) = N$

Optimality criteria

$$\text{A - optimality:} \quad \min_{\xi \in \chi}\{Trace[X'X)^{-1}]\} = \min_{\xi \in \chi}\left\{\sum_{i}^{p} \lambda_i\right\}$$

$$\text{E - optimality:} \quad \min_{\xi \in \chi}\{\text{ largest eigenvalue } [X'X)^{-1}\}$$

Relative Efficiency

$$\text{Eff}(\xi_1;\xi_2) = \left\{ \frac{Det(Cov(\hat{\beta}_{\xi_2}))}{Det(Cov(\hat{\beta}_{\xi_1}))} \right\}^{1/p}$$

% of observations = $(\text{Eff}(\xi_1;\xi_2)^{-1} - 1)100\%$

Analysis of Variance design

y is continuous and x is a nominal group variable

No medication: $D_1 = 0,\ D_2 = 0,\ D_3 = 0$

Only medication A: $D_1 = 1,\ D_2 = 0,\ D_3 = 0$

Only medication B: $D_1 = 0,\ D_2 = 1,\ D_3 = 0$

Both medications A and B: $D_1 = 0,\ D_2 = 0,\ D_3 = 1$

$$y_i = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \varepsilon_i$$

Analysis of Variance design

y = continuous and x is a nominal group variable

No:	$y_i = \beta_0 + \varepsilon_i$
Only A:	$y_i = \beta_0 + \beta_1 + \varepsilon_i$
Only B:	$y_i = \beta_0 + \beta_2 + \varepsilon_i$
A & B:	$y_i = \beta_0 + \beta_3 + \varepsilon_i$

Analysis of Variance design

How can N patients be allocated to four groups?

$n_1 = w_1 N$: No medication

$n_2 = w_2 N$: Only medication A

$n_3 = w_3 N$: Only medication B

$n_4 = w_4 N$: Both medications A and B

$$\sum w_j = 1,$$

A measure for unbalancedness is how different the values of w_j's are.

Vocabulary growth study

Suppose y is continuous, and

SES and Grade are quantitative group variables

		8^{th}	9^{th}	10^{th}	11th
	(1)	n	n	n	n
SES score	(2)	n	n	n	n
	(3)	n	n	n	n

Vocabulary growth study

SES x Grade Level

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

Two hypotheses of interest:

$$H_0: \quad \beta_1 = 0$$

$$H_0: \quad \beta_2 = 0$$

Vocabulary growth study

SES x Grade Level

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

$$\mathrm{Cov}\begin{bmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{bmatrix} = \begin{bmatrix} \dfrac{\sigma_\varepsilon^2}{N(1-r_{12}^2)\mathrm{Var}(x_1)} & -\dfrac{\sigma_\varepsilon^2 \mathrm{Cov}(x_1 x_2)}{N(1-r_{12}^2)\mathrm{Var}(x_1)\mathrm{Var}(x_2)} \\ -\dfrac{\sigma_\varepsilon^2 \mathrm{Cov}(x_1 x_2)}{N(1-r_{12}^2)\mathrm{Var}(x_1)\mathrm{Var}(x_2)} & \dfrac{\sigma_\varepsilon^2}{N(1-r_{12}^2)\mathrm{Var}(x_2)} \end{bmatrix}$$

($k = 3$, but intercept β_0 is discarded)

Vocabulary growth study

SES x Grade Level

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

$\mathrm{Cov}(\hat{\beta})$ decreases when:

- Variance of variables $\mathrm{Var}(x_1)$ and $\mathrm{Var}(x_2)$ increase
- Correlation r_{12} between x_1 and x_2 decreases
- Common error variance σ_ε^2 decreases
- Total number of observation N increases

Vocabulary growth study

D- Optimality criterion

$$\text{Volume} \propto \text{Det}[\text{Var}(\hat{\beta})]$$

Generalized variance:

$$\frac{[\sigma_\varepsilon^2]^2}{N^2(1-r_{12}^2)\text{Var}(x_1)\text{Var}(x_2)}$$

Vocabulary growth study

A- Optimality criterion

Periphery $\propto$ Trace$[Cov(\hat{\beta})]$

Trace criterion:

$$\frac{\sigma_{\varepsilon}^{2}}{N^{2}(1-r_{12}^{2})}[\mathrm{Var}(x_1)^{-1}+\mathrm{Var}(x_2)^{-1}]$$

Vocabulary growth study

Original Design 1

cells	1	2	3	4	5	6	7	8	9	10	11	12
x_1	8	8	8	9	9	9	10	10	10	11	11	11
x_2	1	2	3	1	2	3	1	2	3	1	2	3

$\mathrm{Var}(x_1) = 1.25$ $\quad$ $\mathrm{Var}(x_2) = 0.67$ $\quad$ $r_{12} = 0.0$

$N\mathrm{Var}(\hat{\beta}_1) = 0.80$ $\quad$ $N\mathrm{Var}(\hat{\beta}_2) = 1.50$ $\quad$ $\sigma_\varepsilon^2 = 1.0$

$N\,\mathrm{Trace}[\mathrm{Var}(\hat{\beta})] = 2.30$ $\quad$ $N^2\mathrm{Gen}[\mathrm{Var}(\hat{\beta})] = 1.20$

Vocabulary growth study

Design 2 with **max. Var(x_1)**

cells	1	2	3	4	5	6	7	8	9	10	11	12
x_1	8	8	8	8	8	8	11	11	11	11	11	11
x_2	1	2	3	1	2	3	1	2	3	1	2	3

$\mathrm{Var}(x_1) = 2.25$ $\quad \mathrm{Var}(x_2) = 0.67 \quad r_{12} = 0.0$

$N\mathrm{Var}(\hat{\beta}_1) = 0.44$ $\quad N\mathrm{Var}(\hat{\beta}_2) = 1.50 \quad \sigma_\varepsilon^2 = 1.0$

$N\,\mathrm{Trace}[\mathrm{Var}(\hat{\beta})] = 1.94$ $\quad N^2\mathrm{Gen}[\mathrm{Var}(\hat{\beta})] = 0.67$

Vocabulary growth study

Design 3 unbalanced

cells	1	2	3	4	5	6	7	8	9	10	11	12
x_1	8	8	8	8	8	8	11	11	11	11	11	11
x_2	1	1	3	1	2	3	1	2	3	1	3	3

$\mathrm{Var}(x_1) = 2.25$ $\mathrm{Var}(x_2) = 0.85$ $r_{12} = 0.23$

$N\mathrm{Var}(\hat{\beta}_1) = 0.47$ $N\mathrm{Var}(\hat{\beta}_2) = 1.18$ $\sigma_\varepsilon^2 = 1.0$

$N\,\mathrm{Trace}[\mathrm{Var}(\hat{\beta})] = 1.71$ $N^2\mathrm{Gen}[\mathrm{Var}(\hat{\beta})] = 0.55$

Vocabulary growth study

Design 4 Optimal

cells	1	2	3	4	5	6	7	8	9	10	11	12
x_1	8	8	8	8	8	8	11	11	11	11	11	11
x_2	1	1	1	3	3	3	1	1	1	3	3	3

$$\text{Var}(x_1) = 2.25 \qquad \text{Var}(x_2) = 1.00 \qquad r_{12} = 0.0$$

$$N\text{Var}(\hat{\beta}_1) = 0.44 \qquad N\text{Var}(\hat{\beta}_2) = 1.00 \qquad \sigma_\varepsilon^2 = 1.0$$

$$N\,\text{Trace}[\text{Var}(\hat{\beta})] = 1.44 \qquad N^2\text{Gen}[\text{Var}(\hat{\beta})] = 0.44$$

Vocabulary growth study

Optimal design

		8^{th}	9^{th}	10^{th}	11th
	(1)	3n			3n
SES score	(2)				
	(3)	3n			3n

Vocabulary growth study

Original design

	8^{th}	9^{th}	10^{th}	11th
(1)	n	n	n	n
SES score (2)	n	n	n	n
(3)	n	n	n	n

Vocabulary growth study

Relative efficiencies

$\mathrm{Eff}(\xi_1;\xi_2)$

Design ξ_2

Design ξ_1

	1	2	3	4
1	1.0	0.744	0.677	0.608
2		1.0	0.908	0.816
3			1.0	0.899
4				1.0

Vocabulary growth study

Relative efficiencies

Eff($\xi_1;\xi_2$)	% of obs.
0.8	25%
0.6	67%
0.4	150%
0.2	400%

END