Statistics 203: Introduction to Regression and Analysis of Variance

Experimental design

Jonathan Taylor

Today

- Experimental design in a (small) nutshell.
- Power.
- Classical designs.
- Criteria of optimality.

Why is design important?

- Entire courses based just on design: only a brief overview today.
- Industrial experiments. Often each trial can be very expensive: imagine modelling crash test data for Jaguar....
- Clinical trials. To rule out unexpected selection bias, it is important to sufficiently randomize the study, perhaps keeping some control over heterogeneity.
- A topic of active research: field has progressed well beyond its origins in agricultural field trials. Modern response surface design problems are used frequently in industry.
- Current examples: CFD (Computational Fluid Dynamics) models take ages to run on a computer and one might want to understand how the model depends on the parameters input into it: try a few values of the parameters and "interpolate" (i.e. run a regression).

What makes a good experiment?

- Is the design subject to unforeseen (selection) bias?
- Is it powerful enough to detect a given effect size?
- Is the precision of a certain number of estimators sufficient? i.e. if we assume we know σ^2, can we get a tight enough confidence interval for β_1, say?
- Multivariate generalizations: D-optimality, V-optimality.

Power in a multivariate setting

- To talk about power in a multivariate setting, one needs to know about non-central χ^2, F, t.
- Non-central χ^2: suppose $Z \sim N(\mu, I) \in \mathbb{R}^k$. Then

$$\|Z\|^2 \sim \chi^2_k(\|\mu\|^2).$$

  and $\|\mu\|^2$ is called the *non-centrality* parameter: 0 corresponds to usual χ^2_k.
- Non-central F: $G_1 \sim \chi^2_{\nu_1}(h), G_2 \sim \chi^2_{\nu_2}(0)$ then

$$F = \frac{G_1/\nu_1}{G_2/\nu_2} \sim F_{\nu_1,\nu_2}(h).$$

- Non-central $t_k(h)$: "square root" of $F_{1,k}(h^2)$.

Power example: one-way ANOVA

Source	SS	df	$E(MS)$
Treatments	$SSTR = \sum_{i=1}^{r} n_i \left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right)^2$	$r-1$	$\sigma^2 + \frac{\sum_{i=1}^{r} n_i \alpha_i^2}{r-1}$
Error	$SSE = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\cdot})^2$	$\sum_{i=1}^{r} n_i - r$	σ^2

- Power for testing $H_0 : \alpha_1 = \cdots = \alpha_r = 0$. Under $H_a : \sum_{i=1}^{r} n_i \alpha_i^2 = h$

$$\frac{MSTR}{MSE} \sim F_{df_{TR}, df_E}(\phi)$$

where

$$\phi = \frac{\sum_i n_i \alpha_i^2}{(r-1)\sigma^2} = \frac{h}{(r-1)\sigma^2}.$$

Therefore, if we reject at level α

$$\text{Power} = 1 - P_{df_{TR}, df_E, \phi}(F_{df_{TR}, df_E, 1-\alpha}).$$

where $P_{\nu_1, \nu_2, \phi}$ is the distribution function of $F_{\nu_1, \nu_2}(\phi)$

Determining sample size: power

- Suppose that $n_i = n$, then the non-centrality parameter

$$\phi > (\max_i \alpha_i - \min_i \alpha_i)^2 \frac{n}{(r-1)\sigma^2} = \tilde{\phi}$$

- As power is monotone increasing in ϕ, it is possible to just specify this "range" of α's relative to σ in the form of $\tilde{\phi}$.
- To determine sample size for a given power $1 - \beta$ and Type I error rate α we would plot `Power` as a function of sample size and take the smallest sample size that yields sufficient power.
- Here is the example.

Scheffé's procedure

- We saw Bonferroni correction for simultaneous inference for residuals.
- Suppose we wanted confidence intervals for *many* contrasts

$$\sum_{j=0}^{p} a_j \beta_j .$$

  For instance, all pairwise differences of main effects in a two-way ANOVA model.
- We can use Bonferroni, but as the number of contrasts grows intervals get wider even though they are based on only p random variables: cannot use this to get *confidence bands*.
- Scheffé's procedure: confidence interval for each contrast is

$$\sum_{j=0}^{p} a_j \widehat{\beta}_j \pm SE\left(\sum_{j=0}^{p} a_j \widehat{\beta}_j\right) \sqrt{pF_{p,n-p,1-\alpha}}.$$

Determining sample size: CIs

- Suppose that there is a set of k contrasts that we wish to estimate and each one has a pre-specified target width w_i.
- Scheffé's procedure at level α tells us that the CI for the j-th contrast of interest

$$\sum_{j=1}^{r} a_{i,j}\mu_j$$

has width

$$W_i = 2\sqrt{\sigma^2 \cdot (rF_{r,(n-1)r,1-\alpha}) \sum_{j=1}^{r} \frac{a_{i,j}^2}{n}}$$

- Can replace $rF_{r,(n-1)r,1-\alpha}$ by Bonferroni correction $t^2_{(n-1)r,1-\alpha/k}$.
- Choose n large enough so that $W_i \leq w_i, 1 \leq i \leq k$.

Classical designs

- Randomized.
- Randomized complete block.
- Nested.
- Repeated measures – will come later in random effects.
- Latin square.
- Factorial / fractional factorial.

Randomized design

- This design controls *selection bias* in the experiment: i.e. by assigning "fitter" people to treatment vs. control, looks like treatment is more effective than it is.
- Given r treatments, and nr subjects, assign subjects to treatment are random.
- Assumes implicitly that all subjects are identical – no "controlling" for variables such as gender, age, etc.
- Reduces to a one-way ANOVA model for the treatment effects.

$$Y_{ij} = \mu. + \alpha_i + \varepsilon_{ij}, 1 \leq i \leq r, 1 \leq j \leq n$$

(with usual constraints)

Randomized block design

- If subjects are heterogeneous then some of the variance σ^2 can be attributed to this heterogeneity.
- One may "block" subjects into n homogeneous groups and randomize the r treatments within each block.
- Reduces to a two-way ANOVA model for the block and treatment effects with no interactions.

$$Y_{ij} = \mu_{..} + \rho_i + \tau_j + \varepsilon_{ij}$$

(with usual constraints)

-

SS	df	$E(MS)$
$SSBL = r\sum_i \left(\overline{Y}_{i.} - \overline{Y}_{..}\right)^2$	$n-1$	$\sigma^2 + r\frac{\sum_i \rho_i^2}{r-1}$
$SSTR = n\sum_j \left(\overline{Y}_{.j} - \overline{Y}_{..}\right)^2$	$r-1$	$\sigma^2 + n\frac{\sum_j \tau_j^2}{r-1}$
$SSBL.TR = r\sum_{i,j} \left(Y_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}_{..}\right)^2$	$(n-1)(r-1)$	σ^2

Nested designs

- Example: suppose we are studying the performance of different schools on standardized tests based on the performance of classes within the schools.
- Each school $1 \leq i \leq a$ has b classes taking the tests, of which each class had a different teacher.
- It is natural to think of "school" effect and "teacher" effect, but the teachers taught only within one school: they are *nested* within schools. (Perhaps better to treat this as random ...)

Nested design: ANOVA table

- Model: like a two-way ANOVA model

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$$

with $1 \leq i \leq a$, $1 \leq j \leq b$, $1 \leq k \leq n$.
(with usual constraints)

- Note the $\beta_{j(i)}$'s are not "shared" across schools and can only be estimated within a given school.

-

SS	df	$E(MS)$
$SSA = bn\sum_i \left(\overline{Y}_{i..} - \overline{Y}_{...}\right)^2$	$a-1$	$\sigma^2 + bn\frac{\sum_i \alpha_i^2}{a-1}$
$SSB(A) = n\sum_{i,j} \left(\overline{Y}_{ij.} - \overline{Y}_{i..}\right)^2$	$a(b-1)$	$\sigma^2 + n\frac{\sum_{i,j} \beta_{j(i)}^2}{a(b-1)}$
$SSE = \sum_{i,j,k} \left(Y_{ijk} - \overline{Y}_{ij.}\right)^2$	$ab(n-1)$	σ^2

Latin square

- r treatments: two blocking variables: each block gets all r treatments
- Example:

	Time		
Subject	A	B	C
	B	C	A
	C	A	B

- Model

$$Y_{ijk} = \mu_{...} + \rho_i + \kappa_j + \tau_k + \varepsilon_{ijk} \qquad 1 \leq i, j, k \leq r$$

but only r^2 observations.

- Similar to two-way ANOVA model with no interactions, one replication per cell.
-

$$\overline{Y}_{i..} = \frac{1}{r}\sum_j Y_{ijk}, \overline{Y}_{.j.} = \frac{1}{r}\sum_i Y_{ijk}, \overline{Y}_{..k} = \frac{1}{r}\sum_{i,j} Y_{ijk}.$$

Latin square ANOVA table

- Predicted values

$$\widehat{Y}_{ijk} = \overline{Y}_{i..} + \overline{Y}_{.j.} + \overline{Y}_{..k} - 2\overline{Y}_{...}$$

-

SS	df	$E(MS)$
$SSROW = r\sum_i \left(\overline{Y}_{i..} - \overline{Y}_{...}\right)^2$	$r-1$	$\sigma^2 + r\frac{\sum_i \rho_i^2}{r-1}$
$SSCOL = r\sum_j \left(\overline{Y}_{.j.} - \overline{Y}_{...}\right)^2$	$r-1$	$\sigma^2 + r\frac{\sum_j \kappa_j^2}{r-1}$
$SSTR = r\sum_k \left(\overline{Y}_{..k} - \overline{Y}_{...}\right)^2$	$r-1$	$\sigma^2 + r\frac{\sum_k \tau_k^2}{r-1}$
$SSRem = \sum_{i,j,k}(Y_{ijk} - \widehat{Y}_{ijk})^2$	$(r-1)(r-2)$	σ^2

2^k factorial designs

- Given k factors with 2 levels each there are 2^k possible combinations. Designs including all levels are called 2^k factorial designs.
- To estimate all of them (with replications) becomes quite "expensive".
- Most of the degrees of freedom goes to estimating higher order interactions (which may be less of interest).
- A 2^{k-f} design is a design that drops some combinations in the interest of "cost", but introduces some confounding to the model.
- For instance if $f = 1$ then the experimenter only uses half of the combinations. This means that it may be impossible to "separate" some low order interactions effects with higher order interactions (i.e. they will be *confounded*)
- Which effects are confounded depends on the *defining relation* of the fractional study. Good defining relations leave as many low order interactions (including main effects) estimable.

Fractional design: example

- Three factors: full design

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \end{pmatrix}$$

Fractional design: example

- Three factors: half design

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \end{pmatrix}$$

Design criteria

- In incomplete fractional designs, it gets very tricky to sort out what is confounded with what.
- Need for general criteria to compare designs.
- D-criterion

$$D = \det(X^t X)$$

- V-criterion: given a collection $\{X_1, \ldots, X_k\}$ of "points of interest" in the predictor space

$$\overline{V} = \frac{1}{k} \sum_{i=1}^{k} X_i (X^t X)^{-1} X_i^t.$$