Tuesday, October 27, 2009
12:46 PM

Q1 10 Q3 10 Q5 15
Q2 15 Q4 10 Q6 30

2101 // 80

01	6	
23	6	22
45	6	455
67	6	66
89	6	888
01	7	0100001
23	7	3223333
45	7	454
67	7	66
89	7	
	8	

$\bar{x}$ 69.8

442 // 45

2	
2	8755
3	
3	9885
4	4210110
4	8

$\bar{x} = 36.9$

Tuesday, October 27, 2009
1:26 PM

Ex J – handout 2 typos so far: x_3 is in ~~g~~ grams not mm.

equation (J2) not labelled

Designed experiments / factorial treatment structure

adv. – because we set x's, we can be more sure that they "cause" the observed change in y.
(see § 9.1.2)

– we can balance the expt by observing well chosen values for x's

§ 9.1 on randomization

e.g. Ex J 3 levels of x_1, x_2, x_3
equal # of runs at ea. = 1 combination.

variables in data set are orthogonal, so interpretation of $\hat{\beta}$ is clearer

i.e. $\hat{\beta}_1$ same in $y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$

or $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$

– we can randomize the assignment of treatments to experimental units
(balanced on unknowns, on average)
(we can sometimes 'blind' subjects, so other confounders are also balanced)

Tuesday, October 27, 2009
1:43 PM

Treatments are very often structured in a factorial design

"x"'s are set at different levels

factors can be quantitative or qualitative

Ex J 3 quant. factors each at 3 levels
in HO $x \rightarrow$ len, amp, load
log( ) log( ) log( )
regression lm

HW 2 : factorial analysis
using −1, 0, 1 etc.

Table 9.3
teaching methods 1 qual. factor A B C D E
levels

Sechrate HO
1 quant. factor 5 levels

Table 8.2 3 factors seat height 2 levels } quant.
tire pressure 2 levels }
dynamo on/off qual

Example 9.6 poisons (Table 8.10)
2 factors "treatment" 4 levels
"poison" 3 levels

For all of these, basic starting point is a linear model.

Ex J $y_i = \beta_0 + \beta_1 \log x_{1i} + \cdots + \varepsilon_i$

SR HO $\boxed{y_{tn} = \alpha + \gamma_t + \varepsilon_{tn}}$

linear structure for $E(y)$; additive errors $\sim (0, \sigma^2)$

parametrization changes

even in Ex G $= y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{10} x_{10i} + \varepsilon_i$

models are all $\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$

fitting is always LS

$$\min_{\underset{\sim}{\beta}} \sum_i (y_i - E(y_i))^2 = SS(\beta)$$

R is using lm in all cases.

makes an $n \times 1$ vector of y's

$n \times p$ matrix X — model.matrix

(even if you ask for aov)

Tuesday, October 27, 2009
1:56 PM

When factors are qualitative, usually summarize the analysis with

1) anova table, F-tests (composite H_0)
e.g. $\gamma_1 = \dots = \gamma_T = 0$

2) group means, e.g. $\bar{y}_{t\cdot}$ + est'd std. errors

3) compare various group means, as relevant for the problem

Ex Teaching methods A B C D E

p. 428 1) $\bar{y}_{A\cdot} - \bar{y}_{B\cdot} \pm 1.96 \cdot \widehat{s.e.}(\bar{y}_{A\cdot} - \bar{y}_{B\cdot})$

(uncorrelated) this suggests $\mu_A - \mu_B = 0$

2) $\bar{y}_U = \frac{1}{2}(\bar{y}_A + \bar{y}_B)$

$\bar{y}_C - \bar{y}_U$: CI for $\mu_C - \frac{1}{2}(\mu_A + \mu_B)$ is consistent w/ data

$\bar{y}_D - \bar{y}_U$ ⋮

$\bar{y}_E - \bar{y}_U$

Ex Seedrate (H0) ← 1 factor quantitative

on p 1.: fit $y_{tj} = \alpha + \gamma_t + \varepsilon_{tj}$

under 2 constraints 1) $\gamma_1 = 0$

2) $\sum_{t=1}^{T} \gamma_t = 0$

factor is only a classification

1) is the default param= for factors that aren't ordered

> barley $ SeedRate = factor(. .)

most usual in software packages → > options("contrasts") # checking what's in place

$contrast

unordered "contr.treatment" ordered "contr.poly" $\gamma_1 = 0$

(*) → > options(contrast = c("contr.sum", "contr.poly")) # $\sum_{t=1}^{T} \gamma_t = 0$

> model.matrix(my.lm)

refit the same model $\hat{\gamma}$'s change

↳ refit the same model $\hat{\gamma}$'s change
anova table is the same.

coef(...lm)

	$\hat{\alpha}$	s.e($\hat{\alpha}$)
	$\hat{\gamma}_1$	⋮
	$\hat{\gamma}_2$	⋮
	$\hat{\gamma}_3$	
	$\hat{\gamma}_4$	
	$\hat{\gamma}_5 = -(\hat{\gamma}_1 + \cdots + \hat{\gamma}_4)$	

$H_0: \gamma_2 = 0$ doesn't make sense

$H_0: \gamma_2 = \gamma_1$ does make sense

"make sense" — depends on the problem

- another lm, insisting on treating sucrate as quantitative

now we have a 3rd possible param =
uses orthogonal polynomials (yet another)

$E(y_{tj}): \alpha + \gamma_t$ $\quad$ $\alpha + \gamma_t$ $\quad$ $\alpha + \beta_1 P_1(x) + \beta_2 P_2(x) + \beta_3 P_3(x)$ $+ \beta_4 P_4(x)$

($P_1(x)$: take vals $-2, -1, 0, 1, 2$)

$\gamma_1 = 0$ $\quad$ $\Sigma \gamma_t = 0$

$(\alpha, \gamma_2 \cdots, \gamma_5)$ $\quad$ $(\alpha, \gamma_1, \cdots, \gamma_4)$ $\quad$ $(\alpha, \beta_1, \cdots, \beta_4)$

X $\quad$ X' $\quad$ X''

```
> SR = as.ordered(SR)
> contrasts(SR)
   .L   .Q   .C   .^4
1 -.63  .53 -.31  .12
2 -.32 -.27  .63 -.49
3  0   -.53  0    .74
4  .32 -.27 -.63 -.49
5  .63  .53  .31  .12
```

If factors are quant., then the comparisons of interest are usually related to regression.

That's why R uses "contr. poly", because it gives coefficients relevant to

- linear relationship betw y & x
- quadratic " " "
- cubic
- quartic " . .

very often of interest See also Ex 9.12

> coef.lm or > summary(my lm)

> anova(... split = list( )) see last pg of Ho

Sometimes, we have both, meaning

- one factor (or more)
- quantitative variable (or more)

HW # 1 $z = \pm 1$

x cont^s

$y_{tj} = \alpha + \gamma_t + \beta x_j + \varepsilon_{tj}$ is one model

(*) → OR $\boxed{\mu + \gamma_t + \beta(x_j - \bar{x})} + \varepsilon_{tj}$ Ex H

& analysis of covariance Ex 9.13

OR (if x was set at 5 equally spaced levels)

$$y_{tj} = \beta_0 + \gamma_t + \beta_1 P_1(x) + \dots + \beta_4 P_4(x) + \varepsilon_{tj}$$

using orthog poly's

OR = ... =

§ 9.3.2 for a linear algebra explanation of contrasts

Example 9.6 has 2 factors, 4 obs^s at each comb^n

Example 9.6 has 2 factors, 4 obs per comb of factors

§9.2.4 has the analysis

Tuesday, October 27, 2009
2:49 PM

Example H

$y_i = 1$ if there is a fault in sample i
$\quad\quad 0$ if not

x_i = purity index for sample i

$z_i = -1$ if method = standard
$\quad\quad +1$ if method = modified

$i = 1, \ldots, 44$ 22 batches, divided in half
44 samples

- 1st analysis ignores x_i and uses pairing (paired t-test → binary response)
- regression analysis ignores pairing & pretends 44 indep obs

Not the model $y_i = \alpha + \beta_1 z_i + \beta_2 x_i + \varepsilon_i$

$\varepsilon_i \sim (0, \sigma^2)$

$P_r(y_i = 1) = p_i = \dfrac{e^{\alpha + \beta_1 z_i + \beta_2 x_i}}{1 + e^{\alpha + \beta_1 z_i + \beta_2 x_i}} = E y_i$

random part $y_i = 1$ is p_i ; 0 is $1 - p_i$] binomial dist

$\log\left(\dfrac{p_i}{1-p_i}\right) = \alpha + \beta_1 z_i + \beta_2 x_i$] linear on log-odds

log-odds ratio (works for interpretation)

slight diff. $\log\left(\dfrac{p_i}{1-p_i}\right) = \begin{cases} \alpha + \Delta + \beta(x_i - \bar{x}) & \text{std.} \\ \alpha - \Delta + \beta(x_i - \bar{x}) & \text{mod.} \end{cases}$

$z_i = \begin{matrix} +1 \\ -1 \end{matrix}$ $\begin{matrix} \text{std} \\ \text{mod} \end{matrix}$ Δ instead of β_1