Understanding the analysis of Unbalanced data.

The disussion is in terms of a 2×2 factorial design for simplicity.

		B 1	B 2
A	1	$\bar{y}_{11\cdot}$	$\bar{y}_{12\cdot}$
	2	$\bar{y}_{21\cdot}$	$\bar{y}_{22\cdot}$

balanced $\Leftrightarrow$ same number of experimental units per cel

$y_{ijk} = k^{th}$ observation in cell (i,j), $k = 1, 2, \dots, n$

ANOVA model

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

$$\alpha_1 + \alpha_2 = 0, \quad \beta_1 + \beta_2 = 0$$

$$\gamma_{11} + \gamma_{12} = 0 \quad \gamma_{21} + \gamma_{22} = 0$$

$$\gamma_{11} + \gamma_{21} = 0 \quad \gamma_{12} + \gamma_{22} = 0.$$

ANOVA

SOURCE	df	SS	E(MS)
A	1	SSA	$\sigma^2 + 2n \sum_{i=1}^{2} \alpha_i^2$
B	1	SSB	$\sigma^2 + 2n \sum_{j=1}^{2} \beta_j^2$
AXB	1	SSAB	$\sigma^2 + n \sum_{1}^{2} \sum_{j} \gamma_{ij}^2$
Error	4(n-1)	SSE	σ^2

- To test H_0: No interaction versus H_a: There is an interaction, we use

$$F = \frac{MSAB}{MSE} = \frac{SSAB/1}{SSE/4(n-1)}$$

Reject H_0 if $F > F(\alpha, 1, 4(n-1))$ or if p-value $< \alpha$.

- To test H_0: No A effect versus H_a: there is an A effect, we use

$$F = \frac{MSA}{MSE} = \frac{SSA/1}{SSE/4(n-1)}$$

and reject H_0 if $F > F(\alpha, 1, 4(n-1))$ or if p-value $< \alpha$

- for B we use

$$F = \frac{MSB}{MSE}$$ and reject if

$F > F(\alpha, 1, 4(n-1))$ or if p-value $< \alpha$.

The model can be expressed as a regression model as follows. Let

$$\underset{4n\times 4}{\underset{\sim}{X}} = \begin{array}{c} \begin{array}{cccc} \mu & \alpha_1 & \beta_1 & \gamma_{11} \end{array} \\ \begin{bmatrix} \underset{\sim}{1}_n & \underset{\sim}{1}_n & \underset{\sim}{1}_n & \underset{\sim}{1}_n \\ \underset{\sim}{1}_n & \underset{\sim}{1}_n & -\underset{\sim}{1}_n & -\underset{\sim}{1}_n \\ \underset{\sim}{1}_n & -\underset{\sim}{1}_n & \underset{\sim}{1}_n & -\underset{\sim}{1}_n \\ \underset{\sim}{1}_n & -\underset{\sim}{1}_n & -\underset{\sim}{1}_n & \underset{\sim}{1}_n \end{bmatrix} \end{array} \quad \text{where } \underset{\sim}{1}_n = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

Let $\underset{\sim}{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \gamma_{11} \end{pmatrix}$.

our model can be expressed

$$\underset{\sim}{Y} = \underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

and we can use what we learned in regression to do ~~discuss~~ the test discussed before

For example, to test for interaction, we fit the regression model and just look at the p-value of the t-test corresponding to γ_{11}.

Question: what to if A or B has more than two level and want to use regression approach to test for the presence of the interaction?

Answer: Use Partial F-test

- The beauty of the balanced designs is that the columns of the design matrix are orthogonal.

Result: Suppose we have a regression of p-predictors $X_1, X_2, \dots, X_p$. Suppose These predictors are orthogonal, that is,

$$X_i^T X_j = \sum_{\ell=1}^{n} X_{i\ell} X_{j\ell} = 0 \text{ , then}$$

$$SSR(X_1, \dots, X_p) = SSR(X_1) + SSR(X_2) + \dots + SSR(X_p)$$
$$= SSR(X_1, \dots, X_q) + SSR(X_{q+1}, \dots, X_p)$$

so to test $H_0: \beta_{q+1} = \cdots \beta_{p} = 0$,

we use the F-test

$$F = \frac{[SSR(X_{q+1}) + \cdots + S(X_p)]/(p-q)}{MSE}$$

and reject H_0 if $F > F(\cancel{\alpha}; p-q, n-p-1)$

If the colums are not orthognal, this technique does not work. In the design example I discussed

$$E(SSA) = \sigma^2 + \sum_{i=1}^{4} \sum_{j=1}^{4} m_{ij}^{(A)} \beta_i \beta_j$$

where
$$\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \gamma_{11} \end{pmatrix}$$

$m_{ij}^{(A)}$'s are some constant.

so $\sum_{i=1}^{4} \sum_{j=1}^{4} m_{ij}^{(A)} \beta_i \beta_j = 0 \not\Longrightarrow \alpha_1 = \alpha_2 = 0$ necessarily

- Type I sums of squares

<u>suppose</u> $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i$

$SSR(X_1, \ldots, X_p)$ = regression sum of squares when we fit the model with all the Xs.

we can decompose $SSR(X_1, \ldots, X_p)$ as follows

$$SSR(X_1, \ldots, X_p) = SSR(X_1) + SSR(X_2 \mid X_1) + SSR(X_3 \mid X_1, X_2) + \cdots + SSR(X_p \mid X_1, \ldots, X_{p-1})$$

here $SSR(X_j \mid X_1, \ldots, X_{j-1})$ = additional amount of varibility in the y values that we can explain by adding X_j to a model that contains $X_1, \ldots, X_{j-1}$.

We can also express $SSR(X_1, \ldots, X_p)$ as

$$SSR(X_1, \ldots, X_p) = SSR(X_1, \ldots, X_q) + SSR(X_{q+1}, \ldots, X_p \mid X_1, \ldots, X_q)$$

Clearly to see if we need $X_{q+1}, \ldots, X_p$ in a model that contains $X_{q+1}, \ldots, X_p$, we can use $SSR(X_{q+1}, \ldots, X_p \mid X_1, \ldots, X_q)$ to construct a test for it

The test is called the partial F test and is given by

$$F = \frac{SSR(X_{q+1}, \ldots, X_p \mid X_1, \ldots, X_q)/(p-q)}{\dfrac{SSE(X_1, \ldots, X_p)}{n-p-1}}$$

where $SSE(X_1, \ldots, X_p)$ = error sum of squares for full model.

We reject H_0 if $F > F(\alpha, p-q, n-p-1)$

Now if $(X_1, \ldots, X_q)$ are orthogonal to $(X_{q+1}, \ldots, X_p)$ then

$$SSR(X_{q+1}, \ldots, X_p \mid X_1, \ldots, X_q) = SSR(X_1, \ldots, X_q)$$

when the design is not balanced, Then

$$SSA \neq SSR(A|B)$$

and $$SSB \neq SSR(B|A)$$

and $SSR(A|B)$ is want we need to test for A in presence of B (similar thing for $SSR(B|A)$)

Type III sums of squares prints for us

$SSR(A|B)$ $SSR(B|A)$ - - - .