

# Review of Lectures 22-34

## Linear Models with Random Effects

- ▶ The first half of the semester concerned models with fixed effects only.
- ▶ In these models there is only one variance component - the error variance,  $\sigma_e^2$
- ▶ In this case F-statistics for testing the significance of effects always used the MSE in the denominator
- ▶ In the second half of the semester we considered models additional variance components induced by the inclusion of random factors.

## Linear Models with Random Effects

- ▶ *Fixed factor*: all levels of interest are observed.  
e.g. sex, race, treatment, variety
- ▶ *Random factor*: a sample of levels is observed (from a large population of potential levels)  
e.g. field plots, subjects, variety
- ▶ The “standard assumptions” are that the observed sample of levels is drawn from a normal population with zero mean and unknown variance (e.g.  $\sigma_\alpha^2$ ,  $\sigma_\pi^2$ ,  $\sigma_{\alpha\beta}^2$ , etc.)
- ▶ The denominator for an F-test concerning a fixed effect must be selected so that its expected mean square matches that of the mean square for the effect under the null hypothesis of no effect.

## Linear Models with Random Effects

Source	SS	DF	MS	EMS
A	SSA	$a - 1$	MSA	$\sigma_e^2 + t\sigma_\omega^2 + tn\theta_A$
T	SST	$t - 1$	MST	$\sigma_e^2 + an\theta_T$
W(A)	SSW(A)	$a(n - 1)$	MSW(A)	$\sigma_e^2 + t\sigma_\omega^2$
AT	SSAT	$(a - 1)(t - 1)$	MSAT	$\sigma_e^2 + n\theta_{AT}$
Error	SSE	$a(n - 1)(t - 1)$	MSE	$\sigma_e^2$
		$ant - 1$		

- ▶ Test for AT interaction using  $F = MSAT/MSE$
- ▶ Test for no main effect of factor T ( $H_0 : \theta_T = 0$ ) using  $F = MST/MSE$
- ▶ Test for no main effect of factor A ( $H_0 : \theta_A = 0$ ) using  $F = MSA/MSW(A)$

## Linear Models with Random Effects

- ▶ Typically formulate the model so that  $\mu$  is the expected response averaged over all levels of all factors (both fixed and random).
- ▶ This is accomplished by using sum constraints for all fixed factors (levels of a fixed model term sum to zero over any subscript).
- ▶ For example, consider a randomized blocks experiment in which each combination of the fixed factors occurs exactly once in each block.

$$Y_{ijk} = \mu + \alpha_i + \tau_j + \alpha\tau_{ij} + B_k + \epsilon_{ijk}$$

$$B_k \sim N(0, \sigma_B^2) \quad \epsilon_{ijk} \sim N(0, \sigma_e^2)$$

- ▶ Then, the sum constraints are

$$\sum_i \alpha_i = 0 \quad \sum_j \tau_j = 0 \quad \sum_i \alpha\tau_{ij} = 0 \quad \sum_j \alpha\tau_{ij} = 0$$

## Linear Models with Random Effects

- ▶ These imply

$$E(Y_{ijk}) = \mu + \alpha_i + \tau_j + \alpha\tau_{ij}$$

$$E(\bar{Y}_{i..k}) = \mu + \alpha_i$$

$$E(\bar{Y}_{.jk}) = \mu + \tau_j$$

$$E(\bar{Y}_{...k}) = \mu$$

- ▶ Note that  $\sum_k B_k \neq 0$ , but  $E(\sum_k B_k) = 0$
- ▶ In this example the three factors are *crossed*, because every combination of levels occurs in the experiment.

## Linear Models with Random Effects

- ▶ Interactions involving a random factor are random.
- ▶ Suppose factor A was random in the previous example. Then

$$\alpha_i \sim N(0, \sigma_\alpha^2) \quad \text{and} \quad \alpha\tau_{ij} \sim N(0, \sigma_{\alpha\tau}^2)$$

- ▶ In this case  $\sum_i \alpha_i \neq 0$  and  $\sum_i \alpha\tau_{ij} \neq 0$ , but  $E(\bar{Y}_{i..k}) = \mu$
- ▶ The (total) variance of an individual response is the sum of all the variance components (because the random effects are all independent)

$$\text{var}(Y_{ijk}) = \sigma_B^2 + \sigma_e^2$$

or

$$\text{var}(Y_{ijk}) = \sigma_\alpha^2 + \sigma_{\alpha\tau}^2 + \sigma_B^2 + \sigma_e^2$$

## Linear Models with Random Effects

- ▶ The variance of a sample mean depends on how many levels of each random factor are being averaged over. Consider the second scenario in the previous example

$$\text{var}(\bar{Y}_{.jk}) = \frac{\sigma_\alpha^2}{a} + \frac{\sigma_{\alpha\tau}^2}{a} + \sigma_B^2 + \frac{\sigma_e^2}{a}$$

but

$$\text{var}(\bar{Y}_{.j.}) = \frac{\sigma_\alpha^2}{a} + \frac{\sigma_{\alpha\tau}^2}{a} + \frac{\sigma_B^2}{b} + \frac{\sigma_e^2}{ab}$$

and

$$\text{var}(\bar{Y}_{...}) = \frac{\sigma_\alpha^2}{a} + \frac{\sigma_{\alpha\tau}^2}{at} + \frac{\sigma_B^2}{b} + \frac{\sigma_e^2}{abt}$$

## Linear Models with Random Effects

- ▶ Estimate variance components by the method of moments; i.e. match the random effects mean squares to their expected values
- ▶ From the ANOVA table considered earlier

$$E(\text{MSE}) = \sigma_e^2 \quad \text{and} \quad E(\text{MSW(A)}) = \sigma_e^2 + t\sigma_w^2$$

imply

$$\hat{\sigma}_e^2 = \text{MSE} \quad \text{and} \quad \hat{\sigma}_w^2 = \frac{1}{t} (\text{MSW(A)} - \text{MSE})$$

- ▶ Estimate standard errors by substituting variance component estimates into variance formulas and taking the square root; e.g.

$$\text{se}(\bar{Y}_{.j.}) = \sqrt{\frac{\hat{\sigma}_\alpha^2}{a} + \frac{\hat{\sigma}_{\alpha\tau}^2}{a} + \frac{\hat{\sigma}_B^2}{b} + \frac{\hat{\sigma}_e^2}{ab}}$$

## Linear Models with Random Effects

- ▶ *Pairwise comparisons*: Unlike in the fixed effects case, it is not always true that the variance of a difference between two treatment means is the sum of the variances. This is because random effects that are shared cancel out when differencing.
- ▶ Consider the factorial experiment with two random factors and one fixed

$$\bar{Y}_{.j.} = \mu + \bar{\alpha}_{.} + \tau_j + \bar{\alpha}\tau_{.j} + \bar{B}_{.} + \bar{e}_{.j.}$$

implies

$$\text{var}(\bar{Y}_{.j.} - \bar{Y}_{.j'.}) = 2 \left( \frac{\sigma_{\alpha\tau}^2}{a} + \frac{\sigma_e^2}{ab} \right)$$

## Linear Models with Random Effects

- ▶ More generally, for the a linear contrast,  $\hat{L} = \sum_i c_i \bar{Y}_i$ .

$$\text{var}(\hat{L}) \neq \sum_i c_i^2 \text{var}(\bar{Y}_i)$$

- ▶ For example, suppose the  $\tau_j$ 's are the levels of a time factor and  $L$  is the standardized linear contrast (SLC) measuring the linear effect of time. Then

$$\text{var}(\hat{L}) = \sum_i c_i^2 \left( \frac{\sigma_{\alpha\tau}^2}{a} + \frac{\sigma_e^2}{ab} \right) = \frac{\sigma_{\alpha\tau}^2}{a} + \frac{\sigma_e^2}{ab}$$

because  $\sum_i c_i^2 = 1$ .

- ▶ Note that the contrast coefficients (1 and -1) in the pairwise comparison are not standardized.

## Linear Models with Random Effects

- ▶ A pair of factors are *crossed* if every combination of factor levels occurs in the experiment.
  - ▶ It is possible to change the level of either factor while keeping the other factor fixed.
  - ▶ We can test for interaction if there is replication; i.e. each combination of factor levels occurs more than once.
- ▶ Factor A is *nested* within factor B if the levels of factor A are different at the different levels of factor B.
  - ▶ We cannot change the level of B without also changing the level of factor A.
  - ▶ Wholeplot is nested within the wholeplot treatment factor in a completely randomized split-plot design.
  - ▶ Subjects are nested within the grouping factor in a two-factor repeated measures design with repeated measures on one factor.

## Linear Models with Random Effects

- ▶ Some specific random effect and mixed effects designs considered in this class
  1. Single random factor
  2. Two crossed random factors (with and without replication)
  3. Randomized blocks design
  4. One and two nested factor designs
  5. Completely randomized split plot design
  6. Randomized blocks split plot design
  7. Single factor repeated measures design
  8. Two factor repeated measures with repeated measures on one factor
  9. Cross-over designs
  10. Balanced incomplete block designs

## Hotelling's one-sample $T^2$ test

- ▶ Several response variables on each sampling unit.
- ▶ **Single sample:** Data is summarized in terms of sample mean vector and sample variance-covariance matrix.

$$\bar{\mathbf{y}} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_p \end{pmatrix} \quad \mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}$$

- ▶ Corresponding population quantities are  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$ .
- ▶ Assuming the sample is from a multivariate normal distribution,  $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , test  $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$  using *Hotelling's one-sample  $T^2$  statistic*

## Hotelling's two-sample $T^2$ test

- ▶ Samples from two multivariate normal populations.
- ▶ Assumptions are that the population variances are the same but their means may be different:  $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$  and  $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$
- ▶ Test  $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$  using *Hotelling's two-sample  $T^2$  statistic*
- ▶ Pooled sample variance-covariance matrix:

$$\mathbf{S} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}$$

- ▶ Comment:  $T^2$  is the maximum squared t-statistic for comparing the two groups over all possible linear combinations of the variables - the best linear combination for discriminating between the two groups.

## MANOVA

- ▶ What if we have samples from more than two (treatment) groups?
- ▶ The total sum of squares and products matrix decomposes into between and within groups SSP matrices

$$\mathbf{T} = \mathbf{B} + \mathbf{W}$$

- ▶ Recall that in the univariate case, the F-test is based on the ratio of between to within SS.
- ▶ The analogous quantities in the multivariate case are the SSP matrices,  $\mathbf{B}$  and  $\mathbf{W}$ .
- ▶ The statistics for testing for homogeneity of group mean vectors in the based on multivariate samples are based on the "eigenvalues" of the matrix  $\mathbf{W}^{-1}\mathbf{B}$ .

- ▶ In general, suppose there are  $p$  response variables and  $g$  groups or populations. We want to test the hypothesis  $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \dots = \boldsymbol{\mu}_g$
- ▶ Single factor MANOVA model formulation:

$$\begin{pmatrix} y_{ij1} \\ y_{ij2} \\ \vdots \\ y_{ijp} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} + \begin{pmatrix} \alpha_{i1} \\ \alpha_{i2} \\ \vdots \\ \alpha_{ip} \end{pmatrix} + \begin{pmatrix} \epsilon_{ij1} \\ \epsilon_{ij2} \\ \vdots \\ \epsilon_{ijp} \end{pmatrix}$$

or

$$\mathbf{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{ij}$$

where  $\boldsymbol{\epsilon}_{ij} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$

- ▶ Consider an experiment with  $t$  repeated measurements on  $n$  subjects in  $a$  groups.
- ▶ The standard (univariate) mixed model for this setting is

$$Y_{ijk} = \text{kth repeated measurement on subject } j \text{ in group } i \\ = \mu + \alpha_i + \pi_{j[i]} + \tau_k + \alpha\tau_{ik} + \epsilon_{ijk}$$

$\alpha_i$  = fixed effect of group  $i$

$\pi_{j[i]}$  = random effect of subject  $j$  in group  $i$

$\tau_k$  = fixed main effect of time  $k$

$\alpha\tau_{ik}$  = group by time interaction effects

$$\pi_{j[i]} \sim N(0, \sigma_\pi^2) \quad \epsilon_{ijk} \sim N(0, \sigma_e^2)$$

- ▶ The univariate analysis is based on an assumption of *compound symmetry* for the variance-covariance matrix of the repeated responses on the subjects. This implies *sphericity* for orthogonal contrasts between the levels of the time factor.
- ▶ The response variance is constant over time, and all pairs of responses are *equicorrelated*.

$$\text{cor}(Y_{ijk}, Y_{ijl}) = \rho = \frac{\sigma_\pi^2}{\sigma_\pi^2 + \sigma_e^2}$$

- ▶ In contrast, the only restriction in a MANOVA analysis is that the variance-covariance structure is the same for all groups.

- ▶ *Discriminant Analysis*: Iris data
- ▶ *Principal Components Analysis*: Math test scores
- ▶ *Factor Analysis*: Decathlon results
- ▶ *Cluster Analysis*
  - ▶ *Hierarchical Clustering*: GRE scores
  - ▶ *K-means Clustering*: Gene expression profiles