Stat 315c: Transposable Data Starting with ANOVA

Art B. Owen

Stanford Statistics

Analysis of variance

- ANOVA is a very old subject. It has a few surprises for us. It anticipates many of the issues we face.
- Named vs anonymous entities correspond closely to fixed vs random effects.
- There are complete and computationally elegant inference solutions, under Gaussian assumptions. No need for asymptotics or simulation.
- But even ANOVA seems to break down for the kind of problems we study here.
- So glm versions of ANOVA are not going to suffice.

Anova

- Predictor variables customarily called factors, corresponding parameters are effects
- Extensive vocabulary for meaning and interpretation of variables
  - Fixed vs random effects
  - Nested vs crossed factors
  - Interactions
  - Control vs noise factors
- We'll see why it matters later. If you ignore the nature of the variation you get wrong answers.
- We'll need the ideas but won't be able to use many of the methods.
- The ANOVA setting is pathologically good

Analysis of variance. Y is yield of potatoes

One way layout

- Model: $Y_{ij} \sim N(\mu_j, \sigma^2)$ $j = 1, \ldots, d$, $i = 1, \ldots, n_j$
- EG: d fertilizers, n_j measurements on j'th one
- No connection between Y_{ij} and $Y_{ij'}$
- Does not fit course topic (Later we say: i is "nested" not "crossed")

Analysis of variance. Y is yield of potatoes

One way layout

- Model: $Y_{ij} \sim N(\mu_j, \sigma^2)$ $j = 1, \ldots, d$, $i = 1, \ldots, n_j$
- EG: d fertilizers, n_j measurements on j'th one
- No connection between Y_{ij} and $Y_{ij'}$
- Does not fit course topic (Later we say: i is "nested" not "crossed")

Randomized blocks are closer

- Fertilizers $j = 1, \ldots, d$ on farms $i = 1, \ldots, n$
- Fertilizers are the variables, farms are the cases

Analysis of variance. Y is yield of potatoes

One way layout

- Model: $Y_{ij} \sim N(\mu_j, \sigma^2)$ $j = 1, \ldots, d$, $i = 1, \ldots, n_j$
- EG: d fertilizers, n_j measurements on j'th one
- No connection between Y_{ij} and $Y_{ij'}$
- Does not fit course topic (Later we say: i is "nested" not "crossed")

Randomized blocks are closer

- Fertilizers $j = 1, \ldots, d$ on farms $i = 1, \ldots, n$
- Fertilizers are the variables, farms are the cases

Two way layout fits our theme

- Fertilizers $j = 1, \ldots, d$ and pesticides $i = 1, \ldots, n$
- Both are variables to study

Random and Fixed Effects

Suppose that a predictor variable (effect) takes k levels

Fixed effect

For a fixed effect, we are interested in learning about those k levels

Random and Fixed Effects

Suppose that a predictor variable (effect) takes k levels

Fixed effect

For a fixed effect, we are interested in learning about those k levels

Random effect

For a random effect, the k levels we got are a sample from a larger population. We want our inferences to apply to that larger population.

Random and Fixed Effects

Suppose that a predictor variable (effect) takes k levels

Fixed effect

For a fixed effect, we are interested in learning about those k levels

Random effect

For a random effect, the k levels we got are a sample from a larger population. We want our inferences to apply to that larger population.

Examples

- A = 10 pain killers (aspirin, tylenol,$\cdots$), and,
  B = 5 patients (Vera, Chuck, $\cdots$, Dave)

Random and Fixed Effects

Suppose that a predictor variable (effect) takes k levels

Fixed effect

For a fixed effect, we are interested in learning about those k levels

Random effect

For a random effect, the k levels we got are a sample from a larger population. We want our inferences to apply to that larger population.

Examples

- A = 10 pain killers (aspirin, tylenol,$\cdots$), and,
  B = 5 patients (Vera, Chuck, $\cdots$, Dave)
  A is fixed, B is random

Random and Fixed Effects

Suppose that a predictor variable (effect) takes k levels

Fixed effect

For a fixed effect, we are interested in learning about those k levels

Random effect

For a random effect, the k levels we got are a sample from a larger population. We want our inferences to apply to that larger population.

Examples

- A = 10 pain killers (aspirin, tylenol,$\cdots$), and,
  B = 5 patients (Vera, Chuck, $\cdots$, Dave)
  A is fixed, B is random
- A = 10 batches of chlorpheniramine and $B = 5$ measurement labs

Random and Fixed Effects

Suppose that a predictor variable (effect) takes k levels

Fixed effect

For a fixed effect, we are interested in learning about those k levels

Random effect

For a random effect, the k levels we got are a sample from a larger population. We want our inferences to apply to that larger population.

Examples

- A $=$ 10 pain killers (aspirin, tylenol,$\cdots$), and,
  B $=$ 5 patients (Vera, Chuck, $\cdots$, Dave)
  A is fixed, B is random
- A $=$ 10 batches of chlorpheniramine and $B = 5$ measurement labs
  A is random, B is random

Edge cases

A $= 10$ drill presses and $B = 5$ operators

Edge cases

A = 10 drill presses and $B = 5$ operators

- A is fixed for inferences about those drill presses
- A is random for drill presses in general, as a 'source of variation'
- Similarly for B

Edge cases

A = 10 drill presses and $B = 5$ operators

- A is fixed for inferences about those drill presses
- A is random for drill presses in general, as a 'source of variation'
- Similarly for B

A = measurement times on pills, months $\{0, 1, 3, 6, 12\}$

Edge cases

A = 10 drill presses and $B = 5$ operators

- A is fixed for inferences about those drill presses
- A is random for drill presses in general, as a 'source of variation'
- Similarly for B

A = measurement times on pills, months $\{0, 1, 3, 6, 12\}$

- Those are fixed times, not sampled
- But mass spectography errors (at each month) are a random effect
- So we get a fixed temporal trend plus a random effect

Edge cases

A = 10 drill presses and $B = 5$ operators

- A is fixed for inferences about those drill presses
- A is random for drill presses in general, as a 'source of variation'
- Similarly for B

A = measurement times on pills, months $\{0, 1, 3, 6, 12\}$

- Those are fixed times, not sampled
- But mass spectography errors (at each month) are a random effect
- So we get a fixed temporal trend plus a random effect

A represents 10 US states selected from 50

Edge cases

A = 10 drill presses and $B = 5$ operators

- A is fixed for inferences about those drill presses
- A is random for drill presses in general, as a 'source of variation'
- Similarly for B

A = measurement times on pills, months $\{0, 1, 3, 6, 12\}$

- Those are fixed times, not sampled
- But mass spectography errors (at each month) are a random effect
- So we get a fixed temporal trend plus a random effect

A represents 10 US states selected from 50

- Between 10 out of 10 (a fixed effect)
- and 10 out of ∞ (a random effect)

Nested and crossed effects

Nesting

- The levels of a nested effect are only defined with respect to the containing effect. Also called 'hierarchical'.
- Eg, ingots $j = 1, \ldots, J_i$ nested within 'heats' of steel $i = 1, \ldots, I$.

Nested and crossed effects

Nesting

- The levels of a nested effect are only defined with respect to the containing effect. Also called 'hierarchical'.
- Eg, ingots $j = 1, \ldots, J_i$ nested within 'heats' of steel $i = 1, \ldots, I$.

Crossing

- Levels of a crossed factor retain their meanings at all levels of another factor
- Eg, flame retardants $i = 1, \ldots, I$ in fabrics $j = 1, \ldots, J$
- For this course: we need at least one crossed pair of factors

Nested and crossed effects

Nesting

- The levels of a nested effect are only defined with respect to the containing effect. Also called 'hierarchical'.
- Eg, ingots $j = 1, \ldots, J_i$ nested within 'heats' of steel $i = 1, \ldots, I$.

Crossing

- Levels of a crossed factor retain their meanings at all levels of another factor
- Eg, flame retardants $i = 1, \ldots, I$ in fabrics $j = 1, \ldots, J$
- For this course: we need at least one crossed pair of factors

Factors A at I levels and B at J levels cross to form an "AB interaction" $A \times B$ at IJ levels.
Factors can be nested and crossed in arbitrarily complex ways.
EG: A crossed with B, both nested within $C \times D$

Puzzlers

1. Can we nest a random effect in a random effect?

Puzzlers

1. Can we nest a random effect in a random effect?
   Yes: students within classes within schools within $\cdots$

Puzzlers

1. Can we nest a random effect in a random effect?
   Yes: students within classes within schools within $\cdots$
2. Can we nest a fixed effect in a fixed effect?

Puzzlers

1. Can we nest a random effect in a random effect?
   Yes: students within classes within schools within $\cdots$
2. Can we nest a fixed effect in a fixed effect?
   Yes: car models within manufacturers

Puzzlers

1. Can we nest a random effect in a random effect?
   Yes: students within classes within schools within $\cdots$
2. Can we nest a fixed effect in a fixed effect?
   Yes: car models within manufacturers
3. Can we nest a random effect in a fixed effect?

Puzzlers

1. Can we nest a random effect in a random effect?
   Yes: students within classes within schools within $\cdots$
2. Can we nest a fixed effect in a fixed effect?
   Yes: car models within manufacturers
3. Can we nest a random effect in a fixed effect?
   Yes: movies within studios

Puzzlers

1. Can we nest a random effect in a random effect?
   Yes: students within classes within schools within $\cdots$
2. Can we nest a fixed effect in a fixed effect?
   Yes: car models within manufacturers
3. Can we nest a random effect in a fixed effect?
   Yes: movies within studios
4. Can we nest a fixed effect in a random one?

Puzzlers

1. Can we nest a random effect in a random effect?
   Yes: students within classes within schools within $\cdots$
2. Can we nest a fixed effect in a fixed effect?
   Yes: car models within manufacturers
3. Can we nest a random effect in a fixed effect?
   Yes: movies within studios
4. Can we nest a fixed effect in a random one?
   No. [3 out of 4 isn't bad!]

ANOVA puzzlers ctd

Simple problems fit nesting and crossing paradigms.
But common settings stretch or break them.

Edge cases

- Is 'Department' nested in 'University' or crossed?

ANOVA puzzlers ctd

Simple problems fit nesting and crossing paradigms.
But common settings stretch or break them.

Edge cases

- Is 'Department' nested in 'University' or crossed?
  Nested: j'th dept in uni i unrelated to j'th in uni i'

ANOVA puzzlers ctd

Simple problems fit nesting and crossing paradigms.
But common settings stretch or break them.

Edge cases

- Is 'Department' nested in 'University' or crossed?
  Nested: j'th dept in uni i unrelated to j'th in uni i'
  Crossed: Math/Stat/CS depts in uni's i and i' relate

ANOVA puzzlers ctd

Simple problems fit nesting and crossing paradigms.
But common settings stretch or break them.

Edge cases

- Is 'Department' nested in 'University' or crossed?
  Nested: j'th dept in uni i unrelated to j'th in uni i'
  Crossed: Math/Stat/CS depts in uni's i and i' relate
  Oops: Uni i has no math dept [No problem...just missing data]

ANOVA puzzlers ctd

Simple problems fit nesting and crossing paradigms.
But common settings stretch or break them.

Edge cases

- Is 'Department' nested in 'University' or crossed?
  Nested: j'th dept in uni i unrelated to j'th in uni i'
  Crossed: Math/Stat/CS depts in uni's i and i' relate
  Oops: Uni i has no math dept [No problem...just missing data]
  Oops again: Uni i' has merged it's stat and math depts

ANOVA puzzlers ctd

Simple problems fit nesting and crossing paradigms.
But common settings stretch or break them.

Edge cases

- Is 'Department' nested in 'University' or crossed?
  Nested: j'th dept in uni i unrelated to j'th in uni i'
  Crossed: Math/Stat/CS depts in uni's i and i' relate
  Oops: Uni i has no math dept [No problem...just missing data]
  Oops again: Uni i' has merged it's stat and math depts
- Temperature has levels 45°, 50°, 55°
  Heat time has levels 0, 3, 6, 9 hours
  Get $3 \times 4 = 12$ names for 10 distinct treatments

ANOVA puzzlers ctd

Simple problems fit nesting and crossing paradigms.
But common settings stretch or break them.

Edge cases

- Is 'Department' nested in 'University' or crossed?
  Nested: j'th dept in uni i unrelated to j'th in uni i'
  Crossed: Math/Stat/CS depts in uni's i and i' relate
  Oops: Uni i has no math dept [No problem...just missing data]
  Oops again: Uni i' has merged it's stat and math depts
- Temperature has levels 45°, 50°, 55°
  Heat time has levels 0, 3, 6, 9 hours
  Get $3 \times 4 = 12$ names for 10 distinct treatments
- Are scalar values (like temperature) fixed or random?

ANOVA puzzlers ctd

Simple problems fit nesting and crossing paradigms.
But common settings stretch or break them.

Edge cases

- Is 'Department' nested in 'University' or crossed?
  Nested: j'th dept in uni i unrelated to j'th in uni i'
  Crossed: Math/Stat/CS depts in uni's i and i' relate
  Oops: Uni i has no math dept [No problem...just missing data]
  Oops again: Uni i' has merged it's stat and math depts
- Temperature has levels 45°, 50°, 55°
  Heat time has levels 0, 3, 6, 9 hours
  Get $3 \times 4 = 12$ names for 10 distinct treatments
- Are scalar values (like temperature) fixed or random?
  Usually fixed with comparatively simple generalizations to, eg 51.5°

ANOVA puzzlers ctd

Simple problems fit nesting and crossing paradigms.
But common settings stretch or break them.

Edge cases

- Is 'Department' nested in 'University' or crossed?
  Nested: j'th dept in uni i unrelated to j'th in uni i'
  Crossed: Math/Stat/CS depts in uni's i and i' relate
  Oops: Uni i has no math dept [No problem...just missing data]
  Oops again: Uni i' has merged it's stat and math depts
- Temperature has levels 45°, 50°, 55°
  Heat time has levels 0, 3, 6, 9 hours
  Get $3 \times 4 = 12$ names for 10 distinct treatments
- Are scalar values (like temperature) fixed or random?
  Usually fixed with comparatively simple generalizations to, eg 51.5°
  Can be random eg time in a stationary setting with slow sampling

Edge cases ctd

Fixed $\bigcup$ random effects.

If a discrete variable takes very many levels some rare and some common, then the common values might be treated as fixed, while the rare ones might be random.
Conceptually we can think that the variable has $k = k_F + k_R$ levels of which k_F are so common and important that we treat them as fixed while the other k_R appear rarely enough that we treat them as random. [We expect $k_F \ll k_R$.]

Examples

- For Amazon.com
  - Harry Potter might be a fixed level.
  - Most other books might be random.
  - A book reseller who buys from Amazon might be a fixed level customer
  - Most other customers might be random levels.

Factor types

Control factor

A factor is a control factor if it corresponds to a decision we control

- Placing our ad on the left vs right of the web page
- Blinking vs non-blinking ad
- Using steel or chalk in our struts

Factor types

Control factor

A factor is a control factor if it corresponds to a decision we control

- Placing our ad on the left vs right of the web page
- Blinking vs non-blinking ad
- Using steel or chalk in our struts

Noise factor

A noise factor corresponds to a decision (ordinarily) out of our control

- Customer using dialup vs high speed cable modem
- Customer driving in Texas summer vs Alaska winter

Usually we can actually control the noise factor in experiments

Factor types

Control factor

A factor is a control factor if it corresponds to a decision we control

- Placing our ad on the left vs right of the web page
- Blinking vs non-blinking ad
- Using steel or chalk in our struts

Noise factor

A noise factor corresponds to a decision (ordinarily) out of our control

- Customer using dialup vs high speed cable modem
- Customer driving in Texas summer vs Alaska winter

Usually we can actually control the noise factor in experiments

Uses

- Robust design: Make a good choice of control at all noise levels
- Personalization: Study control $\times$ noise interaction

Why factor types matter

Ignoring fixed vs random can lead to serious errors.

Contribution (effect) of factor A at level i

Factor Type	Effect	Identifying condition
Fixed	α_i	$\sum_i \alpha_i = 0$
Random	a_i	$a_i \sim N(0, \sigma_A^2)$

Two way ANOVA models with IID replicates

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \quad \text{Fixed}$$
$$Y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + \varepsilon_{ijk} \quad \text{Random}$$
$$Y_{ijk} = \mu + \alpha_i + b_j + (\alpha b)_{ij} + \varepsilon_{ijk} \quad \text{Mixed}$$

Interactions

$\alpha\beta$, or αb, or $a\beta$, or ab

ANOVA estimates

Balanced case: $i = 1, \ldots, I$ $j = 1, \ldots, J$ and $k = 1, \ldots, K$

Estimates and means

$$\hat{\mu} = \bar{Y}_{\bullet\bullet\bullet} \equiv \frac{1}{IJK} \sum_i \sum_j \sum_k Y_{ijk}$$

$$\hat{\alpha}_i = \bar{Y}_{i\bullet\bullet} - \hat{\mu} \equiv \frac{1}{JK} \sum_j \sum_k Y_{ijk} - \bar{Y}_{\bullet\bullet\bullet}$$

$$\hat{\beta}_j = \bar{Y}_{\bullet j\bullet} - \hat{\mu} \quad \text{etc.}$$

$$\widehat{(\alpha\beta)}_{ij} = \bar{Y}_{ij\bullet} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j$$

$$= \bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet j\bullet} + \bar{Y}_{\bullet\bullet\bullet}$$

ANOVA estimates

Sums of squares

$$\mathrm{SS}_A = \sum_i \sum_j \sum_k (\bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet})^2 = JK \sum_i (\bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet})^2$$

$$\mathrm{SS}_B = \sum_i \sum_j \sum_k (\bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet\bullet})^2 = \text{etc.}$$

$$\mathrm{SS}_{AB} = \sum_i \sum_j \sum_k (\bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet j\bullet} + \bar{Y}_{\bullet\bullet\bullet})^2$$

$$\mathrm{SS}_E = \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij\bullet})^2$$

Pythagoras:

$$\mathrm{SS}_T \equiv \sum_i \sum_j \sum_k (\bar{Y}_{ijk} - \bar{Y}_{\bullet\bullet\bullet})^2 = \mathrm{SS}_A + \mathrm{SS}_B + \mathrm{SS}_{AB} + \mathrm{SS}_E.$$

Degrees of freedom

Geometry and projections

- Let Y be a vector in $\mathbb{R}^N$ where $N = I \times J \times K$.
- Each SS is the square norm $\|PY\|^2 = Y'P'PY = Y'PY$
- matrix P projects orthogonally onto a subspace (so $P = P'$ and $PP = P$)

DF is the subspace dimension

- The degrees of freedom of SS is the dimension of that subspace
- $\mathrm{DF}_A = I - 1$, $\mathrm{DF}_B = J - 1$, $\mathrm{DF}_{AB} = (I-1)(J-1)$, $\mathrm{DF}_E = IJ(K-1)$, $\mathrm{DF}_T = IJK - 1$

Mean squares

- $\mathrm{MS}_A = \mathrm{SS}_A/\mathrm{DF}_A$ etc
- Make sense for spherically symmetric noise

ANOVA tables

For two fixed factors A and B we summarize via

Source	Df	SS	MS	F
A	$I-1$	SS_A	MS_A	$\mathrm{MS}_A/\mathrm{MS}_E$
B	$J-1$	SS_B	MS_B	$\mathrm{MS}_B/\mathrm{MS}_E$
AB	$(I-1)(J-1)$	SS_{AB}	MS_{AB}	$\mathrm{MS}_{AB}/\mathrm{MS}_E$
E	$IJ(K-1)$	SS_E	MS_E	
T	$IJK-1$	SS_T		

- SS_A captures 'practical significance' of A
- MS_A and F capture 'statistical significance'
- Compare F to $F_{\mathrm{DF}_A,\mathrm{DF}_E}$ distribution to test $H_0 : \alpha_1 = \cdots = \alpha_I = 0$
- Use MS_E to get confidence statements on $\alpha_i - \alpha_{i'}$ etc

Two crossed random effects

$$Y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + \varepsilon_{ijk}$$

Suppose that

$$\begin{aligned} a_i &\sim N(0, \sigma_A^2) && i = 1, \ldots, I \\ b_j &\sim N(0, \sigma_B^2) && j = 1, \ldots, J \\ (ab)_{ij} &\sim N(0, \sigma_{AB}^2) \\ \varepsilon_{ijk} &\sim N(0, \sigma_E^2) && k = 1, \ldots, K \end{aligned}$$

All independent

Expected mean squares

After some gory math

$$
\begin{aligned}
E(\mathrm{MS}_A) &= \sigma^2_E + K\sigma^2_{AB} + JK\sigma^2_A \\
E(\mathrm{MS}_B) &= \sigma^2_E + K\sigma^2_{AB} + IK\sigma^2_B \\
E(\mathrm{MS}_{AB}) &= \sigma^2_E + K\sigma^2_{AB} \\
E(\mathrm{MS}_E) &= \sigma^2_E
\end{aligned}
$$

Upshot

- The AB interaction inflates MS_A
- To infer about unsampled i: use $\mathrm{MS}_A/\mathrm{MS}_{AB}$
- Versus $\mathrm{MS}_A/\mathrm{MS}_E$ when for fixed effects

Mixed models

For A fixed and B random

$$E(\mathrm{MS}_A) = \sigma^2_E + K\sigma^2_{AB} + JK\frac{\sum_i \alpha_i^2}{I-1}$$
$$E(\mathrm{MS}_B) = \sigma^2_E + IK\sigma^2_B$$
$$E(\mathrm{MS}_{AB}) = \sigma^2_E + K\sigma^2_{AB}$$
$$E(\mathrm{MS}_E) = \sigma^2_E$$

'Other effect' rule

- Test fixed effect A via $\mathrm{MS}_A/\mathrm{MS}_{AB}$
- Test random effect B via $\mathrm{MS}_B/\mathrm{MS}_E$

Subtle: $\begin{pmatrix}(\alpha b)_{i1}\\(\alpha b)_{i2}\\\vdots\\(\alpha b)_{iJ}\end{pmatrix} \sim N\left(0, \sigma^2_{AB}\left(I_J - \frac{11'}{J}\right)\right)$ so $\sum_j (\alpha b)_{ij} = 0$

Intuition

True vs nominal sample size

- $I = 5$ drugs (fixed)
- $J = 8$ patients (random)
- $K = 100$ blood pressure measurements per patient per drug
- Now let $K \to \infty$
- Get 5×8 matrix of true $\mu_{ij} \equiv E(Y_{ijk} \mid \text{Drug i \& Patient j})$.
- For comparing drugs, we still just have 8 patients (40 obs)
- Plain regression on all $40K$ obs would use MS_E

Intuition

True vs nominal sample size

- $I = 5$ drugs (fixed)
- $J = 8$ patients (random)
- $K = 100$ blood pressure measurements per patient per drug
- Now let $K \to \infty$
- Get 5×8 matrix of true $\mu_{ij} \equiv E(Y_{ijk} \mid$ Drug i & Patient j).
- For comparing drugs, we still just have 8 patients (40 obs)
- Plain regression on all $40K$ obs would use MS_E

Is $40K$ a true sample size for patients?

Intuition

True vs nominal sample size

- $I = 5$ drugs (fixed)
- $J = 8$ patients (random)
- $K = 100$ blood pressure measurements per patient per drug
- Now let $K \to \infty$
- Get 5×8 matrix of true $\mu_{ij} \equiv E(Y_{ijk} \mid \text{Drug i \& Patient j})$.
- For comparing drugs, we still just have 8 patients (40 obs)
- Plain regression on all $40K$ obs would use MS_E

Is $40K$ a true sample size for patients?

- Maybe, and yes for interactions: from known μ_{ij}'s we could be 100% sure that $\sigma^2_{AB} > 0$.

Pigeonhole model

Operation

- Rectangular table has R rows and C columns
- Each of RC pigeonholes (cells) has N numbers in it
- We sample r rows and c columns
- If row i and col j are sampled, so is pigeonhole ij

  We sample n of the N numbers from pigeonhole ij

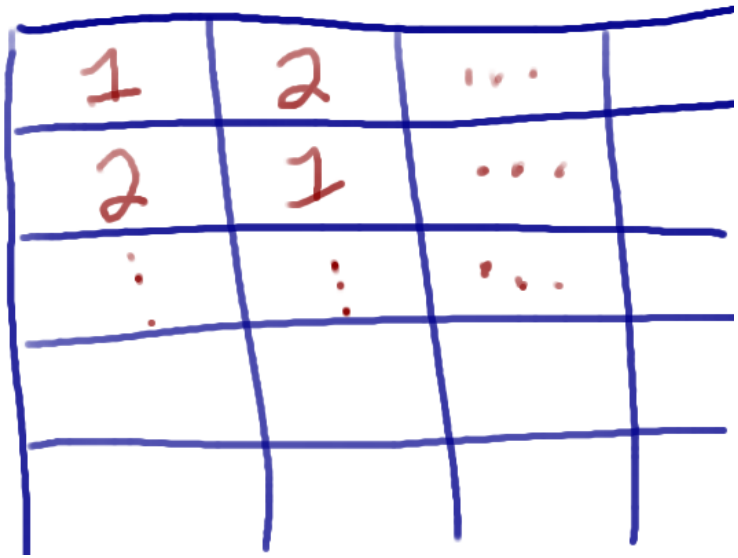
Features

- Due to Cornfield and Tukey
- Needs no assumption of normality or constant variance
- Let $N \to \infty$ to sample more generally
- Take $r = R$ for fixed effects
- Send $R \to \infty$ for random effects

Pigeonhole generality

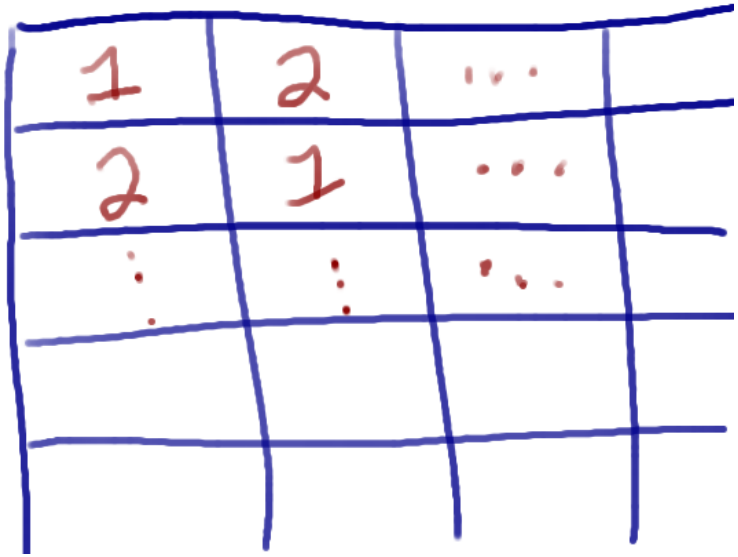
$N(\mu_{ij}, 1)$ cells

With μ_{ij} pictured
Random effects can't handle it
Pigeonhole can



$N(\mu + a_i + b_j + (ab)_{ij}, \sigma^2_{ij})$ cells

With σ^2_{ij} pictured
Random effects can't handle it
Pigeonhole can



Pigeonhole continued

Expected mean squares

Rows	$(1-\frac{n}{N})\sigma_E^2 + n(1-\frac{c}{C})\sigma_{RC}^2 + nc\,\sigma_R^2$
Columns	$(1-\frac{n}{N})\sigma_E^2 + n(1-\frac{r}{R})\sigma_{RC}^2 + nr\sigma_C^2$
Interaction	$(1-\frac{n}{N})\sigma_E^2 + n\sigma_{RC}^2$
Error	σ_E^2

Specialize

- Toggle c/C and r/R to 0 or 1
- Take $N = \infty$
- Recover usual expected mean squares
- ...and more (eg edge case $r < R < \infty$)

Generalize

- extends to $R \times C \times S \times \cdots \times Z$ tables
- is the basis for Anova tests

Large unbalanced random effects

Setting (eg raters i and rated items j)

$$Y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + \varepsilon_{ijk}$$
$$k = 1, \ldots, n_{ij}$$

Goals

- Compare σ^2_A, σ^2_B, σ^2_{AB}, σ^2_E
- Estimate some specific a_i's or b_j's or $(ab)_{ij}$'s

Sparsity

- Most $n_{ij} = 0$
- Most other $n_{ij} = 1$
- So lets just use $\varepsilon_{ij} \equiv (ab)_{ij} + \varepsilon_{ij1}$

Shrinkage estimates

Model and notation

- Now $Y_{ij} = \mu + a_i + b_j + \varepsilon_{ij}$
- Let $n_{i\bullet} = \sum_j n_{ij} = \#$obs for row i, $n_{\bullet j} = \sum_i n_{ij} = \#$obs for col j

Shrinkage

- Given μ, σ_A^2, σ_B^2, $\sigma_E^2 = \mathrm{Var}(\varepsilon_{ij})$
- Put $\bar{Y}_{i\bullet} = \sum_{j(i)} Y_{ij}/n_{i\bullet}$
- Let $\hat{a}_i = \lambda_i(\bar{Y}_{i\bullet} - \mu)$
- Pick λ_i to min $E((a_i - \hat{a}_i)^2)$

Ideally

- $\bar{Y}_{i\bullet} \sim \left(a_i, \frac{\sigma_B^2+\sigma_E^2}{n_{i\bullet}}\right)$ given a_i
- Then take $\lambda_i = \dfrac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma_B^2+\sigma_E^2}{n_{i\bullet}}} = \dfrac{1}{1+\frac{1}{n_{i\bullet}}\frac{\sigma_B^2+\sigma_E^2}{\sigma_A^2}}$

Estimating σ_A^2, σ_B^2, σ_E^2

Eg Netflix data

- 100,000,000 ratings should be enough to pin down μ, σ_A, σ_B and σ_E
- Almost an oracle (for those params)

Methods

1. Moments
2. Maximum likelihood
3. REML

Method of moments

Outline

1. Work out $E(\sum_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2)$ as lin comb of σ_A^2, σ_B^2, σ_E^2
2. Get two more linear combinations, and solve

$$\begin{pmatrix} \mathrm{SS}_1 \\ \mathrm{SS}_2 \\ \mathrm{SS}_3 \end{pmatrix} = \begin{pmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{pmatrix} \begin{pmatrix} \sigma_A^2 \\ \sigma_B^2 \\ \sigma_E^2 \end{pmatrix}$$

Issues

- Sums of squares must be 'free of fixed effects'
- Maybe use $\sum_i n_{i\bullet}(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2$ instead
- And/or replace $\bar{Y}_{\bullet\bullet}$ by $I^{-1}\sum_i \bar{Y}_{i\bullet}$
- We could generate more equations than unknowns
- Usual choice based on variance
- But ... lack of fit is more important

For Netflix data

Estimates

$$\hat{\mu} = 3.604$$

$$\hat{\sigma}^2_{\text{movi}} = 0.272 \qquad \hat{a}_{\text{movi}} = \frac{\bar{Y}_{\text{movi}}}{1 + 5.01/n_{\text{movi}}}$$

$$\hat{\sigma}^2_{\text{cust}} = 0.185 \qquad \hat{b}_{\text{cust}} = \frac{\bar{Y}_{\text{cust}}}{1 + 7.83/n_{\text{cust}}}$$

$$\hat{\sigma}^2_E = 1.178$$

But answer depends on

1. Moment method used
2. Data subset applied to

Note how large $\hat{\sigma}^2_E$ is. That's partly because the model is so simple. Also: should we account for selection bias?

Maximum likelihood and REML

These are the most recommended methods

Model for $y \in \mathbb{R}^N$

$$y = X\beta + Zu + e \qquad \text{X fixed} \quad \text{u random} \quad \text{Z 'incidence'}$$

$$= X\beta + \sum_{\ell=1}^{L} Z_\ell u_\ell + e \qquad \text{eg L} = \text{n. rows} + \text{n. cols}$$

$$= X\beta + \sum_{\ell=0}^{L} Z_\ell u_\ell, \qquad u_\ell \sim N(0, \sigma_\ell^2 I_{d_\ell})$$

For MLE, solve

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

$$\text{tr}(\hat{V}^{-1}Z_\ell Z_\ell') = (y - X\hat{\beta})'\hat{V}^{-1}Z_\ell Z_\ell'\hat{V}^{-1}(y - X\hat{\beta}), \qquad \text{where,}$$

$$\hat{V} = \sum_{\ell=0}^{L} Z_\ell Z_\ell' \hat{\sigma}_\ell^2 \qquad \text{is } N \times N$$

For more

Searle, Casella, McCulloch

- consider 5 moment methods
  - Yule I and II [Raw direct moments]
  - Henderson I, II, and III [BLUE and BLUP]
- REML is
  - MLE based on $K'y \sim N(0, K'VK)$
  - where $K'X\beta = 0$
  - it fixes up $(1 - 1/m)$ like terms
- ML and REML estimation is nasty for large unbalanced data
  - Accounting for mixed effects is hard
  - Even EM looks hard

Bootstrap methods

Here's what I'd do.

Fixed $\times$ fixed

- Treat as regression and resample residuals
- or use 'wild bootstrap' [Essentially $\pm\hat{\varepsilon}_{ij}$]
- out of luck for saturated model
- might then resample unbalancedly (only for saturated where we're desperate)
- Desperate $\cap$ null model $\cdots$ permute rows and/or columns

Random $\times$ fixed

- Resample the random factor
- Problematic if random factor has only few levels
- (We're stuck then anyhow)

Bootstrap methods ctd

Random $\times$ random, McCullagh (2000)

- No consistent bootstrap variance exists for $\hat{\mu} = \frac{1}{IJ}\sum_i \sum_j Y_{ij}$
- But ... see Section 4.6

Pigeonhole bootstrap

- resample rows
- resample cols
- retain intersected cells

Model based bootstrap

- fit $a_i \sim \hat{F}_A$ and $b_j \sim \hat{F}_B$ and $\varepsilon_{ij} \sim \hat{F}_E$
- Take $\hat{Y}_{ij}^{*b} = \hat{\mu} + a_i^{*b} + b_j^{*b} + \varepsilon_{ij}^{*b}$

Near accuracy

Actual variance of $\hat{\mu}$ is

$$\frac{\sigma_A^2}{m} + \frac{\sigma_B^2}{n} + \frac{\sigma_E^2}{mn}$$

Expected bootstrap variance (for pigeon boot or model boot)

$$\sigma_A^2\Big(\frac{m-1}{m^2}\Big) + \sigma_B^2\Big(\frac{n-1}{n^2}\Big) + \sigma_E^2\Big(\frac{3}{mn} - \frac{2}{mn^2} - \frac{2}{m^2n} + \frac{1}{m^2n^2}\Big)$$

Upshot

- Trouble if $\sigma_A^2 = \sigma_B^2 = 0$
- Pretty good if m and n are both large and σ_E^2 not relatively enormous
- This case was balanced

Naive bootstrap

McCullagh's Boot-I

- We have N triples $(i, j, Y_{ij}) \in \mathcal{I} \times \mathcal{J} \times \mathbb{R}$
- Resample them with replacment

Recall Actual variance of $\hat{\mu}$:

$$\frac{\sigma_A^2}{m} + \frac{\sigma_B^2}{n} + \frac{\sigma_E^2}{mn}$$

Expected naive bootstrap variance of $\hat{\mu}$ is

$$\sigma_A^2\Big(\frac{m-1}{m^2n}\Big) + \sigma_B^2\Big(\frac{n-1}{n^2m}\Big) + \sigma_E^2\frac{mn-1}{m^2n^2}$$

Upshot .. it's way too small

- Here we'd need $\sigma_A^2 = \sigma_B^2 = 0$
- What if we're after more than just $\hat{\mu}$?

Sparsely sampled data

Naive bootstrap

- Actual variance of $\hat{\mu} = (1/N)\sum_{ij} Y_{ij}$

$$\sigma_A^2 \frac{1}{N^2}\sum_i n_i^2 + \sigma_B^2 \frac{1}{N^2}\sum_j n_j^2 + \sigma_E^2 \frac{1}{N} \geq \frac{1}{N}\Big(\sigma_A^2 + \sigma_B^2 + \sigma_E^2\Big)$$

- Expected $N/(N-1)\times$ bootstrap variance of $\hat{\mu} = (1/N)\sum_{ij} Y_{ij}$

$$\frac{1}{N}\Big(\sigma_A^2+\sigma_B^2+\sigma_E^2\Big) - \frac{\sigma_A^2}{N(N-1)}\sum_i n_i(n_i-1) - \frac{\sigma_B^2}{N(N-1)}\sum_j n_j(n_j-1).$$

Trouble in proportion to lumpiness:

- Ok when $\max_i n_i = \max_j n_j = 1$
- Bad when some n_i or n_j are huge
- Balanced case not necessarily the worst!

Sparsely sampled data

Pigeonhole bootstrap

- Sample sizes too random on unbalanced data
- Possible fixes: weighted sampling, oversampling

Properties of PBS

- Will sometimes give too little data (left out Harry Potter)
- Sometimes too much (saw HP 3 times)
- Random n_i^*, IE not conditional on sample pattern
- Treats 2 resampled Harry Potters as two different books

Model based bootstrap

- Keeps n_i and n_j fixed
- Requires estimates $\hat{F}_A$, $\hat{F}_B$, $\hat{F}_E$
- Makes strong independence assumptions e.g. $n_i \perp V(Y_{ij} \mid i)$

ANOVA References

1. Box, Hunter and Hunter "Statistics for Experimenters"
   Intuitive intro DOE text
2. D.C. Montgomery "Design and Analysis of Experiments"
   Comprehensive intro DOE text
3. Searle, Casella and McCulloch "Variance Components"
   Extensive coverage of balanced Gaussian random effects
4. Cornfield and Tukey (Article in course web site)
   Presents the pigeonhole model.
5. McCullagh (Article in course web site)
   Perhaps the only one to bootstrap crossed random effects

Structured interaction models

Plain unstructured model

- has $I \times J$ parameters $(\alpha\beta)_{ij}$
- for what may be least interesting term
- and no generalizing structure

Outer product models

- Tukey (1949) 1 df for non-additivity

$$E(Y_{ij}) = \mu + \alpha_i + \beta_j + \lambda\,\alpha_i\beta_j$$

  adds parameter $\lambda \in \mathbb{R}$

- Fisher and MacKenzie (1923) bilinear term

$$E(Y_{ij}) = \mu + \alpha_i + \beta_j + \lambda\,\gamma_i\delta_j$$

  adds parameters $\lambda \in \mathbb{R}$ γ_i and δ_j

Structured interaction models

Plain unstructured model

- has $I \times J$ parameters $(\alpha\beta)_{ij}$
- for what may be least interesting term
- and no generalizing structure

Outer product models

- Tukey (1949) 1 df for non-additivity

$$E(Y_{ij}) = \mu + \alpha_i + \beta_j + \lambda\,\alpha_i\beta_j$$

  adds parameter $\lambda \in \mathbb{R}$

- Fisher and MacKenzie (1923) bilinear term

$$E(Y_{ij}) = \mu + \alpha_i + \beta_j + \lambda\,\gamma_i\delta_j$$

  adds parameters $\lambda \in \mathbb{R}$ γ_i and δ_j much more later