Stat 321: Matrix valued 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 are almost fixed vs random effects.
- There are complete and computationally elegant inference solutions, under Gaussian assumptions. No need for asymptotics or simulation.
- But ANOVA also breaks down for the kind of problems we study here. [Large scale and unbalanced.]
- 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 need the ideas $\cdots$ but can't use many of the methods.
- The ANOVA setting is pathologically good

Analysis of variance. Y is yield of potatoes

One way layout

- Model: $Y_{ij} \sim N(\mu_j, \sigma^2)$ $j = 1, \dots, d$, $i = 1, \dots, n_j$
- 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: $Y_{ij} \sim N(\mu_j, \sigma^2)$ $j = 1, \dots, d$, $i = 1, \dots, n_j$
- 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 $j = 1, \dots, d$ on farms $i = 1, \dots, n$
- Fertilizers are the variables, farms are the cases

Analysis of variance. Y is yield of potatoes

One way layout

- Model: $Y_{ij} \sim N(\mu_j, \sigma^2)$ $j = 1, \dots, d$, $i = 1, \dots, n_j$
- 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 $j = 1, \dots, d$ on farms $i = 1, \dots, n$
- Fertilizers are the variables, farms are the cases

Two way layout fits our theme

- Fertilizers $j = 1, \dots, d$ and pesticides $i = 1, \dots, n$
- 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

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 $j = 1, \dots, J_i$ nested within 'heats' of steel $i = 1, \dots, I$.

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 $j = 1, \dots, J_i$ nested within 'heats' of steel $i = 1, \dots, I$.

Crossing

- Levels of a crossed factor retain their meanings at all levels of another factor
- Eg, flame retardants $i = 1, \dots, I$ in fabrics $j = 1, \dots, J$
- 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 $j = 1, \dots, J_i$ nested within 'heats' of steel $i = 1, \dots, I$.

Crossing

- Levels of a crossed factor retain their meanings at all levels of another factor
- Eg, flame retardants $i = 1, \dots, I$ in fabrics $j = 1, \dots, J$
- 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!]

Head vs long tail

Uneven sampling

- There are often just a few common levels and a great many rare levels.
- This is roughly described by Zipf laws: i'th most popular has $\propto i^{-a}$ events $a \in (1, \infty)$.
- The head has well known entities, the tail is a mishmash

Head vs long tail

Uneven sampling

- There are often just a few common levels and a great many rare levels.
- This is roughly described by Zipf laws: i'th most popular has $\propto i^{-a}$ events $a \in (1, \infty)$.
- The head has well known entities, the tail is a mishmash

E.g. at Amazon.com

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

Head vs long tail

Uneven sampling

- There are often just a few common levels and a great many rare levels.
- This is roughly described by Zipf laws: i'th most popular has $\propto i^{-a}$ events $a \in (1, \infty)$.
- The head has well known entities, the tail is a mishmash

E.g. at Amazon.com

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

Similarly: queries, IP addresses, URLs, phone numbers $\cdots$

More about factors

Control factor

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

- Ad on left/right of page, blinking vs not, etc.
- Using steel or aluminum in auto part

More about factors

Control factor

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

- Ad on left/right of page, blinking vs not, etc.
- Using steel or aluminum in auto part

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

More about factors

Control factor

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

- Ad on left/right of page, blinking vs not, etc.
- Using steel or aluminum in auto part

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.
- You can underestimate the real sampling uncertainty.
- Big errors come from treating random as fixed.
- That is what most regression code does as default.

Large unbalanced random effects

Setting (eg raters i and rated items j)

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

Goals

- Compare σ_A^2, σ_B^2, σ_{AB}^2, σ_E^2
- 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 let's just use $\varepsilon_{ij} \equiv (ab)_{ij} + \varepsilon_{ij1}$ (roll interaction into error)

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

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

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, but they don't scale

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

$$
\begin{aligned}
y &= X\beta + Zu + e \qquad \text{X fixed} \quad \text{u random} \quad \text{Z 'incidence'}\\
&= X\beta + \sum_{\ell=1}^{L} Z_\ell u_\ell + e \qquad \text{eg L} = \text{n. rows} + \text{n. cols}\\
&= X\beta + \sum_{\ell=0}^{L} Z_\ell u_\ell, \qquad u_\ell \sim N(0, \sigma_\ell^2 I_{d_\ell})
\end{aligned}
$$

For MLE, solve

$$
\begin{aligned}
X'\hat{V}^{-1}X\hat{\beta} &= X'\hat{V}^{-1}y\\
\mathrm{tr}(\hat{V}^{-1}Z_\ell Z_\ell') &= (y - X\hat{\beta})'\hat{V}^{-1}Z_\ell Z_\ell'\hat{V}^{-1}(y - X\hat{\beta}), \qquad \text{where,}\\
\hat{V} &= \sum_{\ell=0}^{L} Z_\ell Z_\ell' \hat{\sigma}_\ell^2 \qquad \text{is } N \times N
\end{aligned}
$$

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

These method won't work on big unbalanced data. So it becomes a research issue to get equally good results in a practical way.

Bootstrap methods

Here's what I'd do.

Fixed × 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 × fixed

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

Bootstrap methods ctd

Random × 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)

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

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

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

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 $\hat\mu = (1/N)\sum_{ij} Y_{ij}$

$$\sigma_A^2 \frac{1}{N^2}\sum_i n_i^2 + \sigma_B^2 \frac{1}{N^2}\sum_j n_j^2 + \sigma_E^2 \frac{1}{N} \ge \frac{1}{N}\Bigl(\sigma_A^2 + \sigma_B^2 + \sigma_E^2\Bigr)$$

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

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

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