Statistics 203: Introduction to Regression and Analysis of Variance

Fixed vs. Random Effects

Jonathan Taylor

Today's class

- Random effects.
- One-way random effects ANOVA.
- Two-way mixed & random effects ANOVA.
- Sattherwaite's procedure.

Two-way ANOVA

- Second generalization: more than one grouping variable.
- Two-way ANOVA model: observations: $(Y_{ijk}), 1 \leq i \leq r, 1 \leq j \leq m, 1 \leq k \leq n_{ij}$: r groups in fi rst grouping variable, m groups ins second and n_{ij} samples in (i,j)-“cell”:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \qquad \varepsilon_{ijk} \sim N(0, \sigma^2).$$

- Constraints:
  - $\sum_{i=1}^r \alpha_i = 0$
  - $\sum_{j=1}^m \beta_j = 0$
  - $\sum_{j=1}^m (\alpha\beta)_{ij} = 0, 1 \leq i \leq r$
  - $\sum_{i=1}^r (\alpha\beta)_{ij} = 0, 1 \leq j \leq m.$

Random vs. fixed effects

- In ANOVA examples we have seen so far, the categorical variables are well-defi ned categories: below average fi tness, long duration, etc.
- In some designs, the categorical variable is "subject".
- Simplest example: repeated measures, where more than one (identical) measurement is taken on the same individual.
- In this case, the "group" effect α_i is best thought of as random because we only sample a subset of the entire population of subjects.

When to use random effects?

- A "group" effect is random if we can think of the levels we observe in that group to be samples from a larger population.
- Example: if collecting data from different medical centers, "center" might be thought of as random.
- Example: if surveying students on different campuses, "campus" may be a random effect.

Example: sodium content in beer

- How much sodium is there in North American beer? How much does this vary by brand?
- Observations: for 6 brands of beer, researchers recorded the sodium content of 8 12 ounce bottles.
- Questions of interest: what is the "grand mean" sodium content? How much variability is there from brand to brand?
- "Individuals" in this case are brands, repeated measures are the 8 bottles.

One-way random effects model

- Suppose we take n identical measurements from r subjects.
- $Y_{ij} \sim \mu. + \alpha_i + \varepsilon_{ij}, 1 \leq i \leq r, 1 \leq j \leq n$
- $\varepsilon_{ij} \sim N(0, \sigma^2), 1 \leq i \leq r, 1 \leq j \leq n$
- $\alpha_i \sim N(0, \sigma^2_\mu), 1 \leq i \leq r.$
- We might be interested in the population mean, $\mu.$: CIs, is it zero? etc.
- Alternatively, we might be interested in the variability across subjects, σ^2_μ: CIs, is it zero?

Implications for model

- In random effects model, the observations are no longer independent (even if ε's are independent). In fact

$$\mathrm{Cov}(Y_{ij}, Y_{i'j'}) = \sigma^2_\mu \delta_{i,i'} + \sigma^2 \delta_{j,j'}.$$

- In more complicated mixed effects models, this makes MLE more complicated: not only are there parameters in the mean, but in the covariance as well.
- In ordinary least squares regression, the only parameter to estimate is σ^2 because the covariance matrix is $\sigma^2 I$.

One-way random ANOVA table

Source	SS	df	$E(MS)$
Treatments	$SSTR = \sum_{i=1}^r n\left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right)^2$	$r-1$	$\sigma^2 + n\sigma^2_\mu$
Error	$SSE = \sum_{i=1}^r \sum_{j=1}^n (Y_{ij} - \overline{Y}_{i\cdot})^2$	$(n-1)r$	σ^2

- Only change here is the expectation of $SSTR$ which reflects randomness of α_i's.
- ANOVA table is still useful to setup tests: the same F statistics for fi xed or random will work here.
- Under $H_0 : \sigma^2_\mu = 0$, it is easy to see that

$$\frac{MSTR}{MSE} \sim F_{r-1,(n-1)r}.$$

Inference for $\mu_{\cdot}$

- We know that $E(\overline{Y}_{\cdot\cdot}) = \mu_{\cdot}$, and can show that

$$\mathrm{Var}(\overline{Y}_{\cdot\cdot}) = \frac{n\sigma^2_\mu + \sigma^2}{rn}.$$

- Therefore,

$$\frac{\overline{Y}_{\cdot\cdot} - \mu_{\cdot}}{\sqrt{\frac{SSTR}{(r-1)rn}}} \sim t_{r-1}$$

- Why $r-1$ degrees of freedom? Imagine we could record an infinite number of observations for each individual, so that $\overline{Y}_{i\cdot} \to \mu_i$.
- To learn anything about $\mu_{\cdot}$ we still only have r observations $(\mu_1, \dots, \mu_r)$.
- Sampling more within an individual cannot narrow the CI for $\mu_{\cdot}$.

Estimating σ^2_μ

- From the ANOVA table

$$\sigma^2_\mu = \frac{E(SSTR/(r-1)) - E(SSE/((n-1)r))}{n}.$$

- Natural estimate:

$$S^2_\mu = \frac{SSTR/(r-1) - SSE/((n-1)r)}{n}$$

- Problem: this estimate can be negative! One of the diffi culties in random effects model.

Example: productivity study

- Imagine a study on the productivity of employees in a large manufacturing company.
- Company wants to get an idea of daily productivity, and how it depends on which machine an employee uses.
- Study: take m employees and r machines, having each employee work on each machine for a total of n days.
- As these employees are not *all* employees, and these machines are not *all* machines it makes sense to think of both the effects of machine and employees (and interactions) as random.

Two-way random effects model

- $Y_{ijk} \sim \mu_{\cdot\cdot} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij}, 1 \leq i \leq r, 1 \leq j \leq m, 1 \leq k \leq n$
- $\varepsilon_{ijk} \sim N(0, \sigma^2), 1 \leq i \leq r, 1 \leq j \leq m, 1 \leq k \leq n$
- $\alpha_i \sim N(0, \sigma^2_\alpha), 1 \leq i \leq r.$
- $\beta_j \sim N(0, \sigma^2_\beta), 1 \leq j \leq m.$
- $(\alpha\beta)_{ij} \sim N(0, \sigma^2_{\alpha\beta}), 1 \leq j \leq m, 1 \leq i \leq r.$
- $\text{Cov}(Y_{ijk}, Y_{i'j'k'}) = \delta_{ii'}\sigma^2_\alpha + \delta_{jj'}\sigma^2_\beta + \delta_{ii'}\delta_{jj'}\sigma^2_{\alpha\beta} + \delta_{ii'}\delta_{jj'}\delta_{kk'}\sigma^2.$

ANOVA tables: Two-way (random)

SS	df	$E(SS)$
$SSA = nm\sum_{i=1}^r \left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot}\right)^2$	$r-1$	$\sigma^2 + nm\sigma^2_\alpha + n\sigma^2_{\alpha\beta}$
$SSB = nr\sum_{j=1}^m \left(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdot\cdot\cdot}\right)^2$	$m-1$	$\sigma^2 + nr\sigma^2_\beta + n\sigma^2_{\alpha\beta}$
$SSAB = n\sum_{i=1}^r \sum_{j=1}^m \left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdot\cdot\cdot}\right)^2$	$(m-1)(r-1)$	$\sigma^2 + n\sigma^2_{\alpha\beta}$
$SSE = \sum_{i=1}^r \sum_{j=1}^m \sum_{k=1}^n (Y_{ijk} - \overline{Y}_{ij\cdot})^2$	$(n-1)ab$	σ^2

- To test $H_0 : \sigma^2_\alpha = 0$ use SSA and $SSAB$.
- To test $H_0 : \sigma^2_{\alpha\beta} = 0$ use $SSAB$ and SSE.

Mixed effects model

- In some studies, some factors can be thought of as fixed, others random.
- For instance, we might have a study of the effect of a standard part of the brewing process on sodium levels in the beer example.
- Then, we might think of a model in which we have a fixed effect for "brewing technique" and a random effect for beer.

Two-way mixed effects model

- $Y_{ijk} \sim \mu_{\cdot\cdot} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij}, 1 \leq i \leq r, 1 \leq j \leq m, 1 \leq k \leq n$
- $\varepsilon_{ijk} \sim N(0, \sigma^2), 1 \leq i \leq r, 1 \leq j \leq m, 1 \leq k \leq n$
- $\alpha_i \sim N(0, \sigma^2_\alpha), 1 \leq i \leq r.$
- $\beta_j, 1 \leq j \leq m$ are constants.
- $(\alpha\beta)_{ij} \sim N(0, (m-1)\sigma^2_{\alpha\beta}/m), 1 \leq j \leq m, 1 \leq i \leq r.$
- Constraints:
  - $\sum_{j=1}^m \beta_j = 0$
  - $\sum_{i=1}^r (\alpha\beta)_{ij} = 0, 1 \leq i \leq r.$
  - Cov $\left((\alpha\beta)_{ij}, (\alpha\beta)_{i'j'}\right) = -\sigma^2_{\alpha\beta}/m$
- Cov$(Y_{ijk}, Y_{i'j'k'}) =$
  $\delta_{jj'}\left(\sigma^2_\beta + \delta_{ii'}\frac{m-1}{m}\sigma^2_{\alpha\beta} - (1-\delta_{ii'})\frac{1}{m}\sigma^2_{\alpha\beta} + \delta_{ii'}\delta_{kk'}\sigma^2\right)$

ANOVA tables: Two-way (mixed)

SS	df	$E(MS)$
SSA	$r-1$	$\sigma^2 + nm\sigma^2_\alpha$
SSB	$m-1$	$\sigma^2 + nr\frac{\sum_{j=1}^m \beta_i^2}{m-1} + n\sigma^2_{\alpha\beta}$
$SSAB$	$(m-1)(r-1)$	$\sigma^2 + n\sigma^2_{\alpha\beta}$
$SSE = \sum_{i=1}^r \sum_{j=1}^m \sum_{k=1}^n (Y_{ijk} - \overline{Y}_{ij\cdot})^2$	$(n-1)ab$	σ^2

- To test $H_0 : \sigma^2_\alpha = 0$ use SSA and SSE.
- To test $H_0 : \beta_1 = \cdots = \beta_m = 0$ use SSB and $SSAB$.
- To test $H_0 : \sigma^2_{\alpha\beta}$ use $SSAB$ and SSE.

Confidence intervals for variances

- Consider estimating σ^2_β in the two-way random effects ANOVA. A natural estimate is

$$\widehat{\sigma}^2_\beta = nr(MSB - MSAB).$$

- What about CI?
- A linear combination of χ^2 – but not χ^2.
- To form a confi dence interval for $\widehat{\sigma}^2_\beta$ we need to know distribution of a linear combination of $MS\cdot' s$, at least approximately.

Sattherwaite's procedure

- Given k independent $MS\cdot$'s

$$\widehat{L} \sim \sum_{i=1}^{k} c_i MS_i$$

- Then

$$\frac{df_T \widehat{L}}{\mathbb{E}(\widehat{L})} \text{``}\sim\text{''}\ \chi^2_{df_T}.$$

where

$$df_T = \frac{\left(\sum_{i=1}^k c_i MS_i\right)^2}{\sum_{i=1}^k c_i^2 MS_i^2 / df_i}$$

where df_i are the degrees of freedom of the i-th MS.

- $(1-\alpha)\cdot 100\%$ CI for $\mathbb{E}(\widehat{L})$:

$$L_L = \frac{df_T \times \widehat{L}}{\chi^2_{df_T;1-\alpha/2}}, \qquad L_U = \frac{df_T \times \widehat{L}}{\chi^2_{df_T;\alpha/2}}$$