25 Lecture 25: March 29

Last time

- Lab session

Today

- Announcement: alternative grading path didn't pass (5:5 from last poll + a fail on the first poll)
- Analysis of Variance (JF chapter 8)
  - one-way anova
  - two-way anova

Additional reference

Course notes by Dr. Jason Osborne.

Analysis of Variance

The term analysis of variance is used to describe the partition of the response-variable sum of squares into "explained" and "unexplained" components, noting that this decomposition applies generally to linear models. For historical reasons, analysis of variance (abbreviated ANOVA) also refers to procedures for fitting and testing linear models in which the explanatory variables are categorical.

One-way ANOVA

Suppose that there are *no* quantitative explanatory variables, but only a single factor (categorical data). For example, for a three-category classification, we have the model

$$Y_i = \alpha + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \epsilon_i \tag{1}$$

employing the following coding for the dummy regressors:

Group	D_1	D_2
1	1	0
2	0	1
3	0	0

The expectation of the response variable in each group (i.e. in each category or level of the factor) is the population group mean, denoted by μ_j for the jth group. Equation 1 produces

the following relationship between group means and model parameters:

$$\begin{aligned}
\text{Group 1: } \mu_1 &= \alpha + \gamma_1 \times 1 + \gamma_2 \times 0 = \alpha + \gamma_1 \\
\text{Group 2: } \mu_2 &= \alpha + \gamma_1 \times 0 + \gamma_2 \times 1 = \alpha + \gamma_2 \\
\text{Group 3: } \mu_3 &= \alpha + \gamma_1 \times 0 + \gamma_2 \times 0 = \alpha
\end{aligned}$$

There are three parameters (α, γ_1 and γ_2) and three group means, so we can solve uniquely for the parameters in terms of the group means:

$$\begin{aligned}
\alpha &= \mu_3 \\
\gamma_1 &= \mu_1 - \mu_3 \\
\gamma_2 &= \mu_2 - \mu_3
\end{aligned}$$

Not surprisingly, α represents the mean of the baseline category (Group 3) and that γ_1 and γ_2 captures differences between the other group means and the mean of the baseline category.

notations

Because observations are partitioned according to groups, it is convenient to let Y_{jk} denote the ith observation within the jth of m groups. The number of observations in the jth group is n_j, and the total number of observations is $n = \sum_{j=1}^{m} n_j$. Let $\mu_j \equiv E(Y_{jk})$ be the population mean in group j.

The one-way ANOVA model is

$$Y_{jk} = \mu + \alpha_j + \epsilon_{jk}$$

where μ represents the general level of response variable in the population; α_j represents the effect on the response variable of membership in the jth group; ϵ_{jk} is an error variable that follows the usual linear-model assumptions: $\epsilon_{jk} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

By taking expectations, we have

$$\mu_j = \mu + \alpha_j$$

The parameters of the model are, therefore, underdetermined, for there are $m+1$ parameters (including μ) but only m population group means (recall the dummy variable trap introduced in collinearity). To produce easily interpretable parameters and that estimates and generalizes usefully to more complex models, we impose the sum-to-zero constraint

$$\sum_{j=1}^{m} \alpha_j = 0$$

With the sum-to-zero constraint, we solve for the parameters

$$\begin{aligned}
\hat{\mu} &= \frac{\sum \mu_j}{m} \\
\hat{\alpha}_j &= \mu_j - \mu
\end{aligned}$$

The fitted Y values are the group means for the one-way ANOVA model:

$$\hat{Y}_{jk} = \hat{\mu} + \hat{\alpha}_j$$

and the regression and residual sums of squares therefore take particularly simple forms in one-way ANOVA:

$$RegSS = \sum_{j=1}^{m}\sum_{k=1}^{n_j}(\hat{Y}_{jk} - \bar{Y})^2 = \sum_{j=1}^{m} n_j(\bar{Y}_j - \bar{Y})^2$$
$$RSS = \sum_{j=1}^{m}\sum_{k=1}^{n_j}(Y_{jk} - \hat{Y}_{jk})^2 = \sum_{j=1}^{m}\sum_{k=1}^{n_j}(Y_{jk} - \bar{Y}_j)^2$$

and can be presented in an ANOVA table.

Table 1: General one-way ANOVA table

Source	Sum of Squares	df	Mean Square	F	H_0
Groups	$\sum n_j(\bar{Y}_j - \bar{Y})^2$	$m-1$	$\frac{RegSS}{m-1}$	$\frac{RegMS}{RMS}$	$\alpha_1 = \cdots = \alpha_m = 0$
Residuals	$\sum\sum(Y_{jk} - \bar{Y}_j)^2$	$n-m$	$\frac{RSS}{n-m}$		
Total	$\sum\sum(Y_{jk} - \bar{Y})^2$	$n-1$			

Sometimes, the column of Source can also be denoted with Treatments (for Groups) and Error (for Residuals). And a balanced one-way ANOVA model has the same number of observations in one group (or treatment), in other words, $n_1 = \cdots = n_m = \frac{n}{m}$.

one-way ANOVA example

The following data come from study investigating binding fraction for several antibiotics using $n = 20$ bovine serum samples:

Antibiotic	Binding Percentage	Sample mean
Penicillin G	29.6 24.3 28.5 32.0	28.6
Tetracyclin	27.3 32.6 30.8 34.8	31.4
Streptomycin	5.8 6.2 11.0 8.3	7.8
Erythromycin	21.6 17.4 18.3 19	19.1
Chloramphenicol	29.2 32.8 25.0 24.2	27.8

Question: Are the population means for these 5 treatments plausibly equal?
Answer:

What do we obtain standard errors of parameter estimates? (HW)

Two-Way ANOVA

The inclusion of a second factor permits us to model and test partial relationships, as well as to introduce interactions. Let's take a look at the patterns of relationship that can occur when a quantitative response variable is classified by two factors.

Patterns of Means in the two-way classification

Consider the following table:

	C_1	C_2	$\ldots$	C_c	
R_1	μ_{11}	μ_{12}	$\ldots$	μ_{1c}	$\mu_{1\cdot}$
R_2	μ_{21}	μ_{22}	$\ldots$	μ_{2c}	$\mu_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
R_r	μ_{r1}	μ_{r2}	$\ldots$	μ_{rc}	$\mu_{r\cdot}$
	$\mu_{\cdot 1}$	$\mu_{\cdot 2}$	$\ldots$	$\mu_{\cdot c}$	$\mu_{\cdot\cdot}$

The factors, R and C (for "rows" and "columns" of the table of means), have r and c categories, respectively. The factor categories are denoted R_j and C_k. Within each cell of the design - that is , for each combination of categories $\{R_j, C_k\}$ of the two factors - there is a population cell mean μ_{jk} for the response variable. Extending the dot notation, we have

$$\mu_{j\cdot} \equiv \frac{\sum_{k=1}^{c} \mu_{jk}}{c}$$

is the marginal mean of the response variable in row j.

$$\mu_{\cdot k} \equiv \frac{\sum_{j=1}^{r} \mu_{jk}}{r}$$

is the marginal mean in column k. And

$$\mu_{\cdot\cdot} \equiv \frac{\sum_j \sum_k \mu_{jk}}{r \times c}$$

is the grand mean.

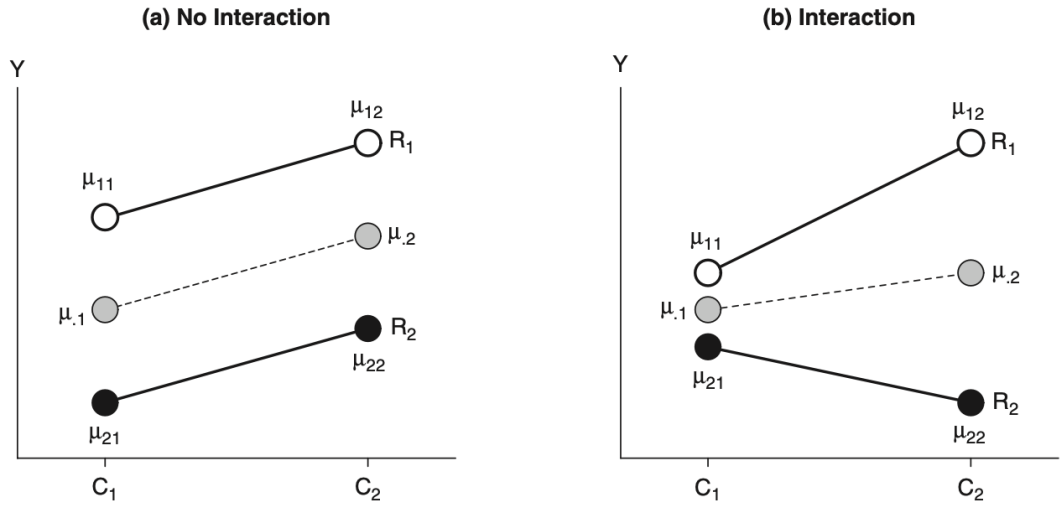

Figure 25.1: Interaction in the two-way classification. In (a), the parallel profiles of means (given by the white and black circles connected by solid lines) indicate that R and C do not interact in affecting Y. The R-effect – that is, the difference between the two profiles – is the same at both C_1 and C_2. Likewise, the C-effect – that is , the rise in the line from C_1 to C_2 – is the same for both profiles. In (b), the R-effect differs at the two categories of C, and the C-effect differs at the two categories of R: R and C interact in affecting Y. In both graphs, the column marginal means $\mu_{\cdot 1}$ and $\mu_{\cdot 2}$ are shown as averages of the cell means in each column (represented by the gray circles connected by broken lines). JF Figure 8.2.