STAT 705 Chapter 16: One-way ANOVA

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 705: Data Analysis II

What is ANOVA?

Analysis of variance (ANOVA) models are regression models with qualitative predictors, called factors or treatments.

Factors have different levels.

For example, the factor "education" may have the levels *high school*, *undergraduate*, *graduate*. The factor "gender" has two levels *female*, *male*.

We may have several factors as predictors, e.g. race and gender may be used to predict annual salary in $.

There are two types of factors:

- Classification (investigator cannot control).
- Experimental (investigator can control).

ANOVA

A control treatment (or control factor level) is sometimes used to measure effects of (new or experimental) treatments under investigation, relative to the "status quo."

e.g. ibuprofin, aspirin, and placebo. We have 3 factor levels. Without placebo, we do not know how iboprofin or aspirin does relative to no pain killer, only relative to each other.

Uses of ANOVA models: find best/worst treatment, measure effectiveness of new treatment, compare treatments.

Often interested in determining whether there is a *difference* in treatments.

Read Sections 16.1–16.8 in the text.

16.3 Cell means model

Have r different treatments or factor levels. At each level i, have n_i observations from group i.

Total number of observations is $n_T = n_1 + n_2 + \cdots + n_r$.

Response is Y_{ij} where $\left\{ \begin{array}{ll} i = 1, \ldots, r & \text{factor level} \\ j = 1, \ldots, n_i & \text{obs. within factor level} \end{array} \right\}$.

Example: Two factors: MS, PhD. Y_{ij} is age in years. Spring of 2014 we observe

$$Y_{11} = 28, Y_{12} = 24, Y_{13} = 24, Y_{14} = 22, Y_{15} = 26, Y_{16} = 23,$$

$$Y_{21} = 29, Y_{22} = 23, Y_{23} = 26, Y_{24} = 25, Y_{25} = 22, Y_{26} = 23, Y_{27} = 38, Y_{28} = 33, Y_{29} = 30, Y_{2,10} = 27.$$

One-way ANOVA model

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2).$$

Can rewrite as

$$Y_{ij} \overset{ind.}{\sim} N(\mu_i, \sigma^2).$$

- Data are normal, data are independent, variance constant across groups.
- μ_i is allowed to be different for each group. $\mu_1, \ldots, \mu_r$ are the r population means of the response. A picture helps.
- Questions: what is $E\{Y_{ij}\}$? What is $\sigma^2\{Y_{ij}\}$?

Matrix formulation

(pp. 683–684, 710–712) For $r = 3$ we have

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2n_2} \\ Y_{31} \\ Y_{32} \\ \vdots \\ Y_{3n_3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \epsilon_{22} \\ \vdots \\ \epsilon_{2n_2} \\ \epsilon_{31} \\ \epsilon_{32} \\ \vdots \\ \epsilon_{3n_3} \end{bmatrix}$$

or

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

16.4 Fitting the model

For $r = 3$, let $Q(\mu_1, \mu_2, \mu_3) = \sum_{i=1}^{3} \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2$.

Need to minumize this over all possible (μ_1, μ_2, μ_3) to find least-squares (LS) solution. Can easily show that $Q(\mu_1, \mu_2, \mu_3)$ has minimum at

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \hat{\mu}_3 \end{bmatrix} = \begin{bmatrix} \bar{Y}_{1\bullet} \\ \bar{Y}_{2\bullet} \\ \bar{Y}_{3\bullet} \end{bmatrix}$$

where $\bar{Y}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$ is the sample mean from the ith group (pp. 687–688).

These $\hat{\boldsymbol{\beta}}$ are also maximum likelihood estimates.

Matrix formula of least-squares estimators ($r = 3$)

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} n_1 & 0 & 0 \\ 0 & n_2 & 0 \\ 0 & 0 & n_3 \end{bmatrix},$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} n_1^{-1} & 0 & 0 \\ 0 & n_2^{-1} & 0 \\ 0 & 0 & n_3^{-1} \end{bmatrix}, \quad \mathbf{X}'\mathbf{Y} = \begin{bmatrix} Y_{1\bullet} \\ Y_{2\bullet} \\ Y_{3\bullet} \end{bmatrix},$$

$$\Rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \bar{Y}_{1\bullet} \\ \bar{Y}_{2\bullet} \\ \bar{Y}_{3\bullet} \end{bmatrix}.$$

Residuals

As in regression (STAT 704),

$$e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \hat{\mu}_i = Y_{ij} - \bar{Y}_{i\bullet}.$$

As usual, $\hat{Y}_{ij}$ is the estimated mean response under the model.

Note that $\sum_{j=1}^{n_i} e_{ij} = 0$. [check this!]

In matrix terms

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Y} - \hat{\mathbf{Y}}.$$

Kenton Food Company Example

- $r = 4$ box designs for a new breakfast cereal.
- 20 stores w/ roughly equal sales volumes picked to participate; $n_i = 5$ is planned for each.
- A fire occurred at one store that had design 3, so ended up with $n_T = 19$ instead of 20, and $n_1 = n_2 = n_4 = 5$ and $n_3 = 4$.

Kenton foods example

```
data kenton;
input sales design @@;
datalines;
 11  1  17  1  16  1  14  1  15  1  12  2  10  2  15  2  19  2  11  2
 23  3  20  3  18  3  17  3  27  4  33  4  22  4  26  4  28  4
;

proc sgscatter;
 plot sales*design;
run;

proc glm plots=all; * zero/one dummy variables, but recover cell means via lsmeans;
 class design;
 model sales=design;
 lsmeans design;
run;
```

16.5 ANOVA table (pp. 690–698)

Define the following

$$Y_{i\bullet} = \sum_{j=1}^{n_i} Y_{ij} = i \text{ group sum},$$

$$\bar{Y}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = i\text{th group mean}$$

$$Y_{\bullet\bullet} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} Y_{ij} = \sum_{i=1}^{r} Y_{i\bullet} = \text{sum all obs.}$$

$$\bar{Y}_{\bullet\bullet} = \frac{1}{n_T} \sum_{i=1}^{r} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n_T} \sum_{i=1}^{r} Y_{i\bullet} = \text{mean all obs.}$$

Sums of squares for treatments, error, and total

$$
\begin{aligned}
\text{SSTO} &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{\bullet\bullet})^2 = \text{variability in } Y_{ij}\text{'s} \\
\text{SSTR} &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\hat{Y}_{ij} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\hat{\mu}_{ij} - \bar{Y}_{\bullet\bullet})^2 \\
&= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{r} n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 \\
&= \text{variability explained by ANOVA model} \\
\text{SSE} &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \hat{Y}_{ij})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i} e_i^2 \\
&= \text{variability NOT explained by ANOVA model}
\end{aligned}
$$

Comments

- As before in regression,

$$\underbrace{\text{SSTO}}_{\text{total}} = \underbrace{\text{SSTR}}_{\text{treatment effects}} + \underbrace{\text{SSE}}_{\text{leftover randomness}}$$

- SSE=0 $\Rightarrow Y_{ij} = Y_{ik}$ for all $j \neq k$
- SSTR=0 $\Rightarrow \bar{Y}_{i\bullet} = \bar{Y}_{\bullet\bullet}$ for $i = 1, \ldots, r$.

ANOVA table (p. 694)

Source	SS	df	MS
SSTR	$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet})^2$	$r-1$	$\text{SSTR}/(r-1)$
SSE	$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet})^2$	n_T-r	$\text{SSE}/(n_T-r)$
SSTO	$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{\bullet\bullet})^2$	n_T-1	

Degrees of freedom

- SSTO has $n_T - 1$ df because there are n_T $Y_{ij} - \bar{Y}_{\bullet\bullet}$ terms in the sum, but they add up to zero (1 constraint).
- SSE has $n_T - r$ df because there are n_T $Y_{ij} - \bar{Y}_{i\bullet}$ terms in the sum, but there are r constraints of the form $\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\bullet}) = 0$.
- SSTR has $r - 1$ df because there are r terms $n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})$ in the sum, but they sum to zero (1 constraint).

Estimated mean squares

$$E\{\text{MSE}\} = \sigma^2, \quad \text{MSE is unbiased estimate of } \sigma^2$$

$$E\{\text{MSTR}\} = \sigma^2 + \frac{\sum_{i=1}^{r} n_i(\mu_i - \mu_\bullet)^2}{r-1},$$

where $\mu_\bullet = \sum_{i=1}^{r} \frac{n_i \mu_i}{n_T}$ is weighted average of $\mu_1, \ldots, \mu_r$ (pp. 696–698).

If $\mu_i = \mu_j$ for all $i, j \in \{1, \ldots, r\}$ then $E\{\text{MSTR}\} = \sigma^2$, otherwise $E\{\text{MSTR}\} > \sigma^2$.

Hence, if any group means are different then $\frac{E\{\text{MSTR}\}}{E\{\text{MSE}\}} > 1$.

16.6 F test of $H_0 : \mu_1 = \cdots = \mu_r$

Fact: If $\mu_1 = \cdots = \mu_r$ then

$$F^* = \frac{\text{MSTR}}{\text{MSE}} \sim F(r-1, n_T - r).$$

To perform α-level test of $H_0 : \mu_1 = \cdots = \mu_r$ vs. H_a : some $\mu_i \neq \mu_j$ for $i \neq j$,

- Accept if $F^* \leq F(1-\alpha, r-1, n_T - r)$ or p-value $\geq \alpha$.
- Reject if $F^* > F(1-\alpha, r-1, n_T - r)$ or p-value $< \alpha$.

p-value $= P\{F(r-1, n_T - 1) \geq F^*\}$.

Example: Kenton Foods

Comments

- If $r = 2$ then $F^* = (t^*)^2$ where t^* is t-statistic from 2-sample pooled-variance t-test.
- The F-test may be obtained from the general nested linear hypotheses approach (big model / little model). Here the full model is $Y_{ij} = \mu_i + \epsilon_{ij}$ and the reduced is $Y_{ij} = \mu + \epsilon_{ij}$.

$$F^* = \frac{\left[\frac{SSE(R) - SSE(F)}{dfE_R - dfE_F}\right]}{\frac{SSE(F)}{dfE_F}} = \frac{MSTR}{MSE}.$$

16.7 Alternative formulations

SAS will fit the cell means model (discussed so far) with a `noint` option in `model` statement; however, the F-test will not be correct. Your textbook discusses an alternative parameterization that is not easy to get out of the SAS procedures we will use.

By default, SAS fits the model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

where $\alpha_r = 0$.

- $E\{Y_{rj}\} = \mu$; μ is the cell-mean for the rth level.
- For $i < r$, $E\{Y_{ij}\} = \mu + \alpha_i$; α_i is i's offset to group r's mean μ.

Note that SAS's default corresponds to a regression model where categorical predictors are modeled using the usual zero-one dummy variables. In class, let's find the design $\mathbf{X}$ for SAS's model for $r = 3$ and $n_1 = n_2 = n_3 = 2$.

SAS's baseline & offset model

Even though SAS parameterizes the model differently, with the rth level as baseline, the ANOVA table and F-test is the same as the cell means model.

Also $\hat{\mu} = \bar{Y}_{r\bullet}$ and $\hat{\alpha}_i = \bar{Y}_{i\bullet} - \bar{Y}_{r\bullet}$ are the OLS and MLE estimators. These are reported in SAS. Use, e.g. `model sales=design / solution;`

The cell means $\hat{\mu}_i$ are obtained in SAS by adding `lsmeans` to `glm` or `glimmix`.