STAT 705 Chapter 16: One-way ANOVA

Adapted from 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. ibuprofen, aspirin, and placebo. We have 3 factor levels. Without the placebo, we do not know how effective ibuprofen or aspirin are relative to no pain killer, only relative to each other.

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

We are 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

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

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

The response is Y_{ij} where

$$\begin{cases} i = 1, \ldots, r & \text{factor level} \\ j = 1, \ldots, n_i & \text{obs. within factor level.} \end{cases}$$

Example: Two factors: MS, PhD. Y_{ij} is age in years. In Spring 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

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

We can rewrite as:

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

- Data are normal, data are independent, the variance is constant across groups.
- μ_i is allowed to be different for each group; the ANOVA model is *nonparametric*.
- Questions: What is $E\{Y_{ij}\}$? What is $\sigma^2\{Y_{ij}\}$?

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

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

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

We need to minimize this over all possible (μ_1, μ_2, μ_3) to find the least-squares (LS) solution. We can easily show that $Q(\mu_1, \mu_2, \mu_3)$ has a minimum at

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

where $\bar{Y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$ 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 ($r = 3$)

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

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

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

Residuals

As in regression (STAT 704),

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

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

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

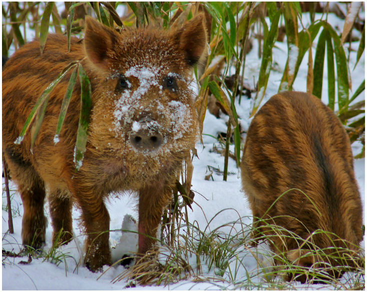
In matrix terms

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

Feral hog rooting activity example

- $r = 4$ habitat types (Bottomland Hardwood, Cypress/Tupelo Slough, Upland Pine, Muck Swamp)
- 20 x 50-meter tracts randomly selected within these habitats. ($n_i \equiv 3$).
- The tracts were monitored on a bi-monthly basis for 18 months; we will consider a single month. One of the Cypress-tupelo tracts was flooded, so $n_T = 11$ rather than 12, and $n_1 = n_3 = n_4 = 3$ and $n_2 = 2$.
- The response will be rooting damage in each of 1000 1 x 1 square meter cells; we will treat it as continuous for this analysis.

Feral Hogs



: Juvenile feral hogs in snowstorm

: Juvenile feral hogs rooting

Bottomland Hardwoods

: Large sweetgum

: Second-growth forest

Cypress-Typelo Sloughs

: Re-sprouted tupelo slough

: Cypress-tupelo slough

Uplands

: Longleaf pine savannah

: Longleaf pine in “rocket” stage

Muck Swamps

: Cinnamon fern at muck swamp edge

: Muck swamp

Feral hog rooting activity example

```
data rooting;
input habitat $ activity @@;
rootroot=sqrt(activity);
datalines;
BLH 139 BLH 228 BLH 275 CTS 45 CTS 127 CTS .
U 0 U 45 U 16  MS 145 MS 124 MS 240
 ;

proc sgplot;
scatter x=habitat y=activity;
run;

proc glm plots=all data=rooting; * zero/one dummy variables, but recover cell means via lsmeans;
 class habitat;
 model rootroot=habitat;
 lsmeans habitat;
run;
```

16.5 ANOVA table (pp. 690–698)

Define the following:

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

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

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

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

Sums of squares for treatments, error, and total

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

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 $j \neq k$ and for $i = 1, \ldots, r$.
- SSTR$=0 \Rightarrow \bar{Y}_{i.} = \bar{Y}_{..}$ for $i = 1, \ldots, r$.

ANOVA table (p. 694)

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

Degrees of freedom

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

Assuming $\mu_1 = \cdots = \mu_r$, Cochran's Theorem (Section 2.7) shows that $SSTR/\sigma^2 \sim \chi^2_{r-1}$ and $SSE/\sigma^2 \sim \chi^2_{n_T - r}$ and they are independent.

Expected mean squares

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

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

where $\mu_\cdot = \sum_{i=1}^{r} \frac{n_i \mu_i}{n_T}$ is the 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 an α-level test of $H_0 : \mu_1 = \cdots = \mu_r$ vs. H_a : some $\mu_i \neq \mu_j$ for $i \neq j$,

- Fail to reject H_0 if $F^* \leq F(1-\alpha, r-1, n_T - r)$ or p-value $\geq \alpha$.
- Reject H_0 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: Feral hog rooting activity

Comments

- If $r = 2$ then $F^* = (t^*)^2$ where t^* is the t-statistic from a 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 obtain from 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 $\hat{\mu} = \bar{Y}_{r\cdot}$ and $\hat{\alpha}_i = \bar{Y}_{i\cdot} - \bar{Y}_{r\cdot}$ 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`.