Analysis of Variance
Bios 662

Michael G. Hudgens, Ph.D.

mhudgens@bios.unc.edu

http://www.bios.unc.edu/∼mhudgens

2008-10-21 13:34

Outline

- Introduction
- Alternative models
- SS decomposition
- Example w/ SAS, R

Analysis of Variance Model

- Ch 10 text (skip 10.3-10.5); Ch 12
- How do we test hypotheses about the mean of more than 2 groups? Analysis of variance (ANOVA) model
- *Definition 10.1* An *analysis of variance model* is a linear regression model in which the predictor variables are classification variables. The categories of a variable are called the *levels* of the variable.
- Categorical predictor variables are also called *qualitative factors*

Notation

- Let Y_{ij} be the j^{th} observation in the i^{th} group
- $i = 1, \ldots, K;\ j = 1, \ldots, n_i$
- Let $N = \sum_{i=1}^{K} n_i$
- $\bar{Y}_{i\cdot} = \sum_j Y_{ij}/n_i$

ANOVA Model and Hypotheses

- Assume $Y_{ij} \sim N(\mu_i, \sigma^2)$
- Want to test

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$$

versus

$$H_A : \text{ at least one } \neq$$

Two variance estimators

- The pooled estimate of σ^2 is:

$$s_p^2 = \frac{\sum_{i=1}^{K}(n_i - 1)s_i^2}{\sum_{i=1}^{K}(n_i - 1)}$$

- Under H_0, the (weighted) variance of the $\bar{Y}_{i\cdot}$'s will estimate σ^2

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{K} n_i(\bar{Y}_{i\cdot} - \bar{Y})^2}{K - 1}$$

where

$$\bar{Y} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i} Y_{ij}}{N}$$

ANOVA: F test

- It can be shown under H_0:

$$(N-K)s_p^2/\sigma^2 \sim \chi^2_{N-K}$$

$$(K-1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{K-1}$$

and s_p^2 and $\hat{\sigma}^2$ are independent

- Therefore, under H_0,

$$F \equiv \frac{\hat{\sigma}^2}{s_p^2} \sim F_{K-1,N-K}$$

ANOVA: F test

- To test H_0,

$$C_\alpha = \{F : F > F_{1-\alpha;K-1,N-K}\}$$

- The test uses $F > F_{1-\alpha;K-1,N-K}$ because under H_A,

$$E(\hat{\sigma}^2) > E(s_p^2)$$

- In particular, $E(s_p^2) = \sigma^2$ whereas

$$E(\hat{\sigma}^2) = \sigma^2 + \frac{\sum_i n_i(\mu_i - \mu)^2}{K-1}$$

  where μ is the overall mean defined in equation (1) below

ANOVA: Example

- Ex: Passive smoking and lung function
- A study was conducted to compare the lung function of groups of smokers and non-smokers. Lung function was measured by forced mid-expiratory flow (FEF)

ANOVA: Example

FEF for males by smoking status

Group	n_i	Mean (L/sec)	sd (L/sec)
Non-smokers	200	3.78	0.79
Passive smokers	200	3.30	0.77
Noninhalers	50	3.32	0.86
Light smk.	200	3.23	0.78
Mod. smk.	200	2.73	0.81
Heavy smk.	200	2.59	0.82

ANOVA: Example

$$C_{.05} = \{F > F_{5,1044;.95} = 2.22\}$$

$$s_p^2 = \frac{199(.79)^2 + 199(.77)^2 + \cdots + 199(.82)^2}{1044} = 0.636$$

$$\hat{\sigma}^2 = \frac{200(3.78 - 3.14)^2 + \cdots + 200(2.59 - 3.14)^2}{5} = 36.875$$

- $F = 36.875/0.636 = 58.0$; Reject H_0
- Reference: NEJM 302(13): 720-3, 1980.

Aside: Obtaining Quantiles/CDFs

- In R

```
> qf(.95,5,1044)
[1] 2.222674

> pf(2.222674,5,1044)
[1] 0.95
```

- In SAS

```
data;
  y = finv(.95,5,1044);
  y1 = quantile('F',.95,5,1044);
  fy = cdf('F',2.22267,5,1044);

proc print;
                 Obs        y           y1          fy

                  1      2.22267     2.22267     0.95000
```

Cell Means Model

- Heretofore, we have looked at ANOVA model

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

for $i = 1, 2, \ldots, K$; $j = 1, 2, \ldots, n_i$ where

$$\epsilon_{ij} \sim N(0, \sigma^2) \text{ for all } i, j$$

Factor Effects Model

- Altenatively, an equivalent model is

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

for $i = 1, 2, \ldots, K$; $j = 1, 2, \ldots, n_i$ where

$$\mu = \frac{\sum_{i=1}^{K} n_i \mu_i}{N} \tag{1}$$

$$\alpha_i = \mu_i - \mu$$

and

$$\epsilon_{ij} \sim N(0, \sigma^2) \text{ for all } i, j$$

Factor Effects Model

- Note typo in text page 363
- Constraint

$$\sum_{i=1}^{K} n_i \alpha_i = 0$$

- α_i does not denote type I error

Model Equivalency

- Equivalency of null hypotheses

$$H_0 : \mu_1 = \cdots = \mu_K \iff H_0 : \alpha_i = 0; i = 1, 2, \ldots, K$$

- α_i is called the i^{th} *main effect* or *factor effect*

$$\begin{aligned} Y_{ij} &= \mu + (\mu_i - \mu) + \epsilon_{ij} \\ &= \mu + \alpha_i + \epsilon_{ij} \\ &= \text{mean} + i^{th} \text{ main effect} + \text{error} \end{aligned}$$

- Data can be partitioned similarly

$$\begin{aligned} Y_{ij} &= \bar{Y} + (\bar{Y}_i. - \bar{Y}) + (Y_{ij} - \bar{Y}_i.) \\ &= \bar{Y} + a_i + e_{ij} \end{aligned}$$

ANOVA: Sum of Squares

- It can be shown (below)

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y})^2 = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(\bar{Y}_{i\cdot}-\bar{Y})^2+\sum_{i=1}^{K}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\cdot})^2$$

- That is

$$\begin{aligned} SST &= SSA + SSW \\ &= (K-1)\hat{\sigma}^2 + (N-K)s_p^2 \end{aligned}$$

ANOVA: Sum of Squares

- Expected value of sum of squares

$$E\{\sum_{i=1}^{K} n_i(\bar{Y}_{i\cdot} - \bar{Y})^2\} = \sum_{i=1}^{K} n_i\alpha_i^2 + (K-1)\sigma^2$$

$$E\{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2\} = (N-K)\sigma^2$$

- Under $H_0 : \alpha_1 = \cdots = \alpha_K = 0$,

$$E\{\sum_{i=1}^{K} n_i(\bar{Y}_{i\cdot} - \bar{Y})^2\} = (K-1)\sigma^2$$

ANOVA: F test

- Therefore, under H_A : at least one $\alpha_i \neq 0$,

$$E(F) > 1$$

- I.e. we reject H_0 if F is too large

$$C_\alpha = \{F : F > F_{1-\alpha;K-1,N-k}\}$$

ANOVA Table

Source of variation	df	MS	F
Among groups	$K-1$	$\hat{\sigma}^2 = \frac{\sum_{i=1}^{K} n_i(\bar{Y}_{i\cdot} - \bar{Y})^2}{K-1}$	MSA/MSW
Within groups	$N-K$	$s_p^2 = \frac{\sum_{i=1}^{K}(n_i-1)s_i^2}{N-K}$	
Total	$N-1$		

ANOVA: Sum of Squares Proof

- Start with

$$\sum_{ij}(Y_{ij} - \overline{Y})^2 = \sum_{ij}(Y_{ij} - \overline{Y}_{i\cdot} + \overline{Y}_{i\cdot} - \overline{Y})^2$$

- RHS equivalent to

$$\sum_{ij}(Y_{ij} - \overline{Y}_{i\cdot})^2 + \sum_{ij}(\overline{Y}_{i\cdot} - \overline{Y})^2 + 2\sum_{ij}(Y_{ij} - \overline{Y}_{i\cdot})(\overline{Y}_{i\cdot} - \overline{Y}).$$

- Last term can be written as

$$2\sum_{i}\{(\overline{Y}_{i\cdot} - \overline{Y})\sum_{j}(Y_{ij} - \overline{Y}_{i\cdot})\},$$

which equals zero since

$$\sum_{j}(Y_{ij} - \overline{Y}_{i\cdot}) = 0$$

for all i.

ANOVA: E(SSW) Proof

$$E(SSW) = E\{\sum_{ij}(Y_{ij} - \overline{Y}_{i\cdot})^2\}$$

$$= E\left\{\sum_i (n_i - 1)\frac{\sum_j (Y_{ij} - \overline{Y}_{i\cdot})^2}{n_i - 1}\right\}$$

$$= \sum_i (n_i - 1)E(s_i^2)$$

$$= \sum_i (n_i - 1)\sigma^2$$

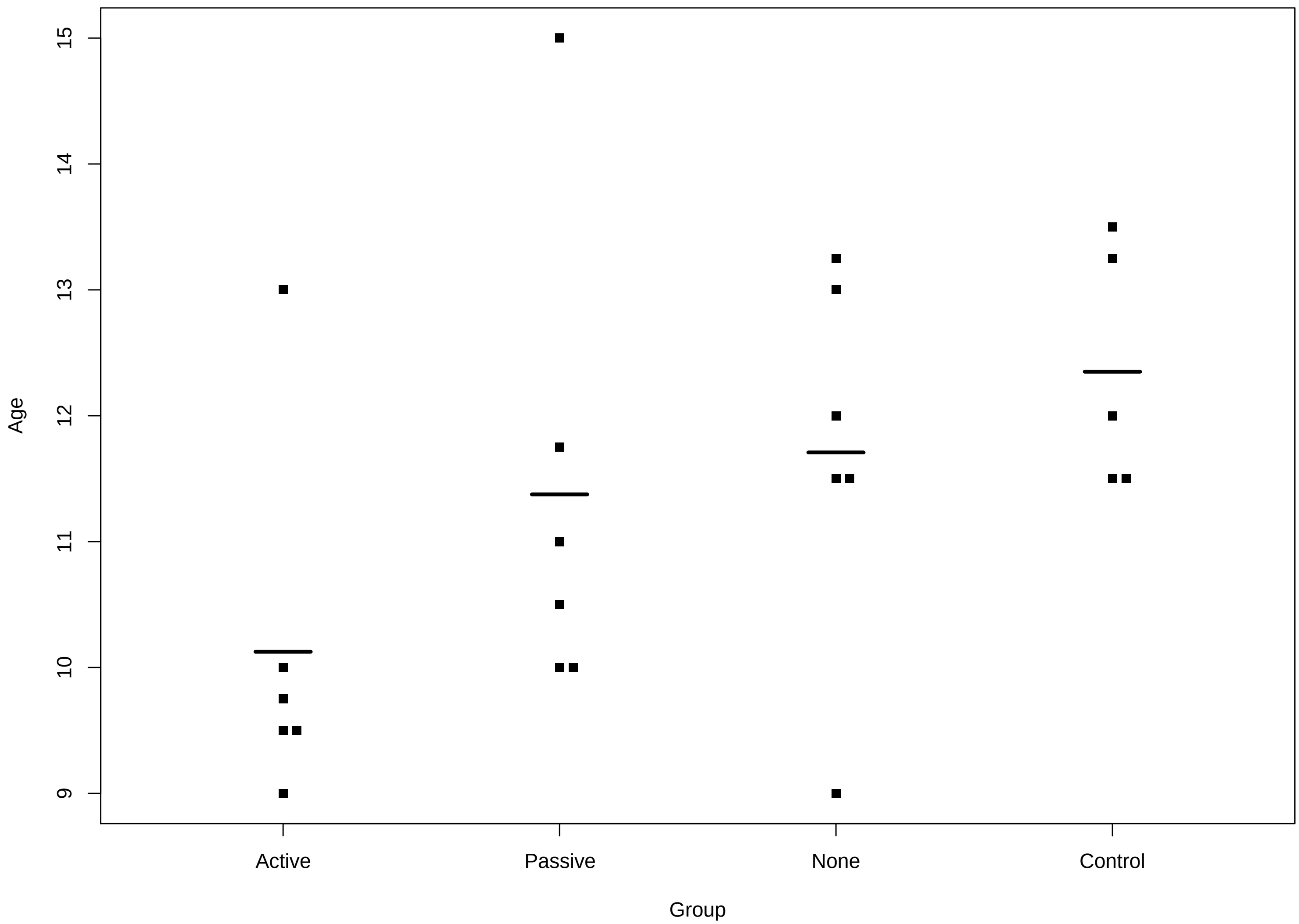
$$= (N - K)\sigma^2$$

ANOVA: Example

- Table 10.1: Distribution of ages (in months) at which infants first walked alone

Active Group	Passive Group	No-Exercise Group	Eight-week Control group
9.00	11.00	11.50	13.25
9.50	10.00	12.00	11.50
9.75	10.00	9.00	12.00
10.00	11.75	11.50	13.50
13.00	10.50	13.25	11.50
9.50	15.00	13.00	

ANOVA: Example



ANOVA: SAS

```
proc glm data=one; * proc anova data=one;
  class group;
  model age=group;
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	14.77780797	4.92593599	2.14	0.1285
Error	19	43.68958333	2.29945175		
Corrected Total	22	58.46739130			

ANOVA: SAS

```
* factor effects model;
data two;
 set one;
 x1=0; x2=0; x3=0;
 if group="active" then x1=1;
 if group="passive" then x2=1;
 if group="no" then x3=1;
 if group="eight" then do; x1=x2=x3=-6/5; end;

proc reg data=two;
 model age = x1  x2  x3;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	14.77781	4.92594	2.14	0.1285
Error	19	43.68958	2.29945		
Corrected Total	22	58.46739			

ANOVA: R

```
> av <- aov(age ~ group)
> anova(av)

Analysis of Variance Table

Response: age
          Df Sum Sq Mean Sq F value Pr(>F)
group      3 14.778   4.926  2.1422 0.1285
Residuals 19 43.690   2.299
```