BIOS 662 Fall 2018

Analysis of Variance, Part I

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

https://sakai.unc.edu/portal

Outline

- Introduction
- Alternative models
- SS decomposition
- Example using SAS, R

Analysis of Variance Model

- Chapter 10 of the text (skip 10.3-10.5); chapter 12
- How do we test hypotheses about the mean of more than two 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 $N = \displaystyle\sum_{i=1}^{K} n_i$
- $\bar{Y}_{i\cdot} = \displaystyle\sum_{j} Y_{ij}/n_i$

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

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

Two Variance Estimators

- The pooled estimator 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 $\bar{Y}_i.$s will estimate σ^2

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

where

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

ANOVA: F Test

- It can be shown that 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,

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

- The test uses $F > F_{K-1,N-K;1-\alpha}$ 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) a few pages ahead

ANOVA: Example

- 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 expiratory flow (FEF)
- 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
Non-inhalers	50	3.32	0.86
Light smokers	200	3.23	0.78
Mod. smokers	200	2.73	0.81
Heavy smokers	200	2.59	0.82

ANOVA: Example cont.

$$C_{0.05} = \{F > F_{5,1044;0.95} = 2.22\}$$

$$s_p^2 = \frac{199(0.79)^2 + 199(0.77)^2 + \cdots + 199(0.82)^2}{1044} = 0.636$$

$$\hat{\sigma}^2 = \frac{200(3.78 - 3.158)^2 + \cdots + 200(2.59 - 3.158)^2}{5} = 36.987$$

- $F = 36.987/0.636 = 58.17 > 2.22$; so reject H_0
- Reference: White JR, Froeb HF. *N Engl J Med* 302(13): 720-3, 1980. (Results presented here may differ from those in the original manuscript because of rounding.)

Aside: Obtaining Quantiles/CDFs

- In R

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

> 1-pf(58.17,5,1044)
[1] 0
```

- In SAS

```
data;
   y = finv(0.95,5,1044);
   y1 = quantile('F',0.95,5,1044);
   fy = cdf('F',58.17,5,1044);
proc print;

Obs        y            y1          fy

 1      2.22267     2.22267          1
```

Cell Means Model

- The version of the ANOVA model we have looked at so far is called the *cell means 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

- An equivalent model is the *factor effects model*

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

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

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

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

and

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

- Note typo in the text on page 363
- Here α_i does not denote type I error

Factor Effects Model

- Constraint: $\sum_{i=1}^{K} n_i \alpha_i = 0$
- Suppose $K = 4$, then from the constraint,

$$n_1\alpha_1 + n_2\alpha_2 + n_3\alpha_3 + n_4\alpha_4 = 0$$

and so

$$\alpha_4 = -(n_1\alpha_1 + n_2\alpha_2 + n_3\alpha_3)/n_4$$

Thus

$$\begin{aligned}
Y_{1j} &= \mu + 1\alpha_1 + \epsilon_{1j} \\
Y_{2j} &= \mu + 1\alpha_2 + \epsilon_{2j} \\
Y_{3j} &= \mu + 1\alpha_3 + \epsilon_{3j} \\
Y_{4j} &= \mu - \frac{n_1}{n_4}\alpha_1 - \frac{n_2}{n_4}\alpha_2 - \frac{n_3}{n_4}\alpha_3 + \epsilon_{4j}
\end{aligned}$$

Model Equivalence

- Equivalence of null hypotheses

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

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

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

- Data can be partitioned similarly

$$\begin{aligned} Y_{ij} &= \bar{Y} + (\bar{Y}_{i\cdot} - \bar{Y}) + (Y_{ij} - \bar{Y}_{i\cdot}) \\ &= \bar{Y} + a_i + e_{ij} \end{aligned}$$

Reference Group Model

- Another equivalent model is the *reference group model*
- One group is chosen as the reference; suppose it is group 1
- Then

$$\begin{aligned} Y_{1j} &= \mu_1 + \epsilon_{1j} \\ Y_{ij} &= \mu_1 + (\mu_i - \mu_1) + \epsilon_{ij}, \quad i = 2, 3, \ldots, K \\ &= \mu_1 + \beta_i + \epsilon_{ij}, \quad i = 2, 3, \ldots, K \end{aligned}$$

for

$$j = 1, 2, \ldots, n_i$$

and

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

- Null hypothesis:

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_K = 0$$

ANOVA: Sum of Squares

- It can be shown (see a few pages ahead) that

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

- That is,

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

- SSW is also referred to as SSE

ANOVA: Sum of Squares

- Expected value of sum of squares

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

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

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

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

ANOVA: F Test and ANOVA Table

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

$$E(F) > 1$$

- That is, we reject H_0 if F is too large

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

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

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

- The RHS is equivalent to

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

- The last term can be written as

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

  which equals zero because

$$\sum_{j}(Y_{ij} - \bar{Y}_{i\cdot}) = 0 \quad \text{for all } i$$

ANOVA: $\boldsymbol{E}$(SSW) Proof

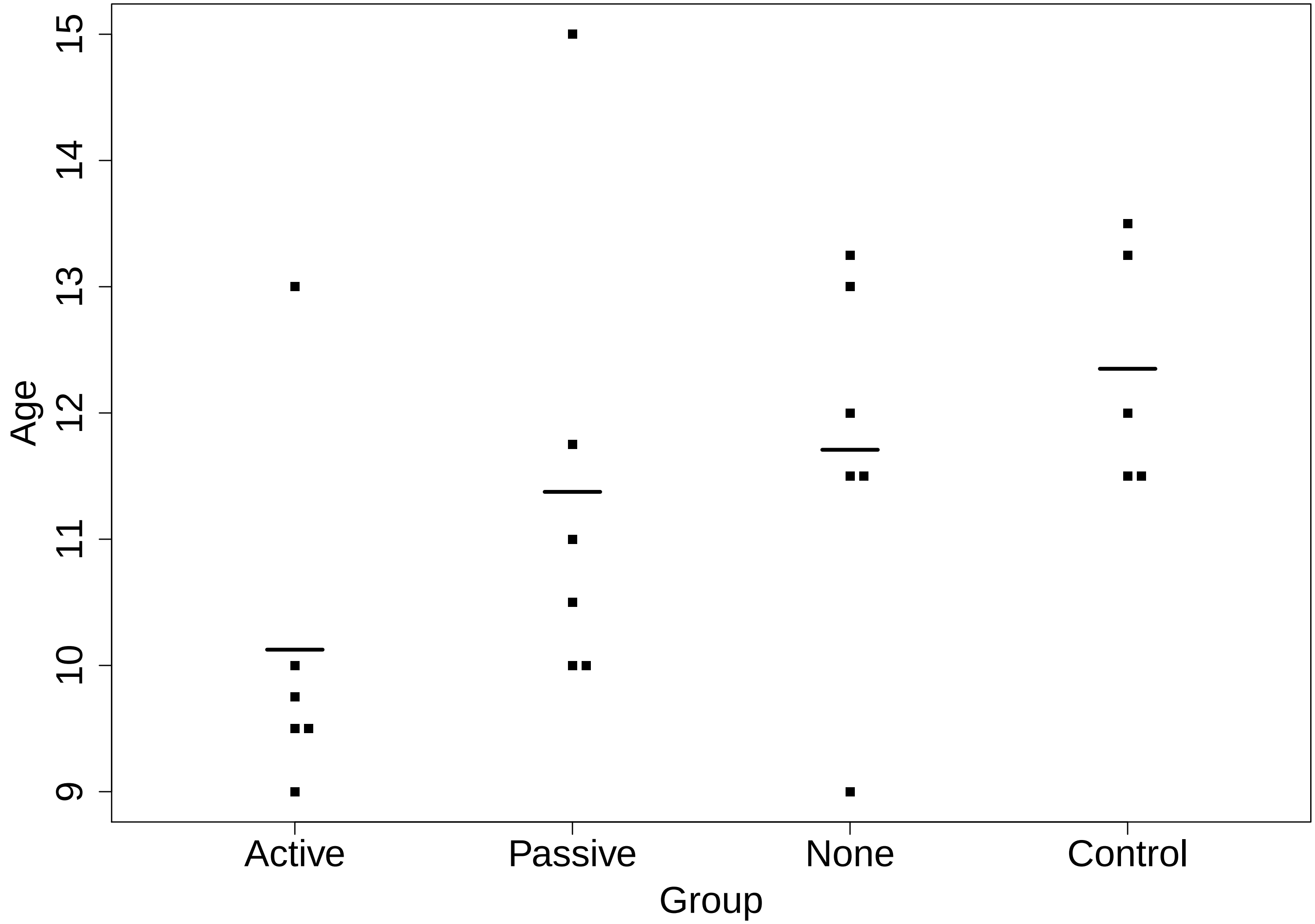
$$
\begin{aligned}
E(\text{SSW}) &= E\Big(\sum_{ij}(Y_{ij} - \bar{Y}_{i\cdot})^2\Big) \\
&= E\left(\sum_i (n_i - 1)\frac{\sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2}{n_i - 1}\right) \\
&= \sum_i (n_i - 1)E(s_i^2) \\
&= \sum_i (n_i - 1)\sigma^2 \\
&= (N - K)\sigma^2
\end{aligned}
$$

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 cont.



ANOVA: SAS – Cell Means Model

```
proc anova data=one;
* Using the following proc statement yields exactly the same ANOVA table;
* proc glm data=one;
   class group;
   model age=group;
```

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;
      else if group="passive" then x2=1;
      else if group="no" then x3=1;
      else if group="eight" then do; x1=x2=x3=-6/5; end;

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

Analysis of Variance

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: SAS – Reference Group Model

```
data three;
   set one;
   x2=0; x3=0; x4=0;
   if group="passive" then x2=1;
      else if group="no" then x3=1;
      else if group="eight" then x4=1;

proc reg data=three;
   model age = x2  x3  x4;
```

Analysis of Variance

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

```
> group <- as.factor(group)
> 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  4.9259  2.1422 0.1285
Residuals 19 43.690  2.2995
```