

# BIOS 7345: Advanced Regression Analysis I

**Andrew J. Spieker, Ph.D.**

Assistant Professor of Biostatistics  
Vanderbilt University Medical Center

Set 7: ANOVA and the coefficient of determination

Version: 09/26/2023

# TABLE OF CONTENTS

- 1 ANOVA: Motivation
- 2 One-way ANOVA
- 3 Two-way ANOVA
- 4 Coefficient of determination

## Ideas:

- We have talked about ANOVA previously (generally as a source of rank-deficient design matrices).
- ANOVA is an abbreviation for **analysis of variance**. We will see in this set of notes why it is so termed.
- Historically, ANOVA has been the approach of choice in the social sciences, laboratory research, etc. - allowing one to compare means across discrete subgroups.

## Ideas:

- Regression: design matrix is front and center:
  - ▶  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ .
- ANOVA: design matrix is an afterthought. Instead, we express the outcome in terms of group-specific means.
- Example:  $i = 1, \dots, n_j$  observations subgroups  $j = 0, 1$ .
  - ▶ Possible model:  $Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$ .
  - ▶ Possible model:  $Y_{ij} = \alpha_j + \epsilon_{ij}$ .
- Example:  $i = 1, \dots, n_{kj}$  observations in subgroups defined by  $j = 1, \dots, J$  and  $k = 1, \dots, K$ .
  - ▶ Possible model:  $Y_{ijk} = \mu_{jk} + \epsilon_{ijk}$ .
  - ▶ Possible model:  $Y_{ijk} = \alpha_j + \beta_k + \epsilon_{ijk}$ .

# TABLE OF CONTENTS

- 1 ANOVA: Motivation
- 2 One-way ANOVA
- 3 Two-way ANOVA
- 4 Coefficient of determination

**Example:** One-way ANOVA with two groups

- Suppose we have  $i = 1, \dots, n_j$  independent observations in subgroups  $j = 1, 2$  (shared outcome variance,  $\sigma^2$ ). Let  $N = n_1 + n_2$ .

	Observations	Mean
Population 1	$y_{11}, \dots, y_{1n_1}$	$\bar{y}_1.$
Population 2	$y_{21}, \dots, y_{2n_2}$	$\bar{y}_2.$
Overall		$\bar{y}_{..}$

- Model:  $Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$ .
- Suppose we seek to test  $H_0 : \alpha_1 = \alpha_2$ . We can use the  $F$ -statistic:

$$\begin{aligned}
 F &= \frac{(\text{RSS}_H - \text{RSS}) / (J - 1)}{\text{RSS} / (N - 2)} \\
 &= \frac{(n_1(\bar{y}_1. - \bar{y}_{..})^2 + n_2(\bar{y}_2. - \bar{y}_{..})^2) / (2 - 1)}{(\sum_{i=1}^{n_1} (y_{1i} - \bar{y}_1.)^2 + \sum_{i=1}^{n_2} (y_{2i} - \bar{y}_2.)^2) / (N - 2)}
 \end{aligned}$$

# ONE-WAY ANOVA

**Example:** One-way ANOVA with  $J$  groups

- Suppose we have  $i = 1, \dots, n_j$  independent observations in subgroups  $j = 0, 1, \dots, J$  (shared outcome variance,  $\sigma^2$ ). Let  $N = \sum_{j=1}^J n_j$ .

	Observations	Mean
Population 1	$y_{11}, \dots, y_{1n_1}$	$\bar{y}_1.$
Population 2	$y_{21}, \dots, y_{2n_2}$	$\bar{y}_2.$
$\vdots$	$\vdots$	$\vdots$
Population $J$	$y_{J1}, \dots, y_{Jn_J}$	$\bar{y}_J.$
Overall		$\bar{y}_{..}$

- Model:  $Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$ .
- To test  $H_0 : \alpha_1 = \dots = \alpha_J$ , we can use the  $F$ -statistic:

$$F = \frac{(\text{RSS}_H - \text{RSS}) / (J - 1)}{\text{RSS} / (N - 2)} = \frac{(\sum_{j=1}^J n_j (\bar{y}_{j.} - \bar{y}_{..})^2) / (J - 1)}{(\sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ji} - \bar{y}_{j.})^2) / (N - J)}$$

**Example:** One-way ANOVA with  $J$  groups

- The  $F$ -statistic has an exact  $F$ -distribution under the null hypothesis *if* the errors are normally distributed.
- If normality does not hold, the  $F$ -distribution is approximate for large samples.



**Example:** One-way ANOVA with  $J$  groups

- Why can the quantity  $RSS_H - RSS$  be written in this format?
  - ▶ Note first that  $RSS_H = \sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ji} - \bar{y}_{..})^2$ .
  - ▶ Then, take note of the following very useful decomposition:

$$\sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ji} - \bar{y}_{..})^2 = \sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ji} - \bar{y}_{j.})^2 + \sum_{j=1}^J n_j (\bar{y}_{j.} - \bar{y}_{..})^2$$

$$SS_T = SS_W + SS_B$$

- The total sum of squares can be written as the within-group sum of squares plus the between-group sum of squares.
- The residual sum of squares is none other than the within-group sum of squares.

**Example:** One-way ANOVA with  $J$  groups

- Suppose  $n_1 = \dots = n_J =: n$ , so that  $N = n \times J$ .
- An ANOVA table typically would have the following information:

Source of variation	df	Sum of squares	Mean sum of squares
Between-group	$N - J = J(n - 1)$	$SS_B$	$SS_B / (N - J)$
Within-group	$J - 1$	$SS_W$	$SS_W / (J - 1)$
Total	$N - 1 = nJ - 1$	$SS_T$	$SS_T / (N - 1)$

- With this, we can see more readily why this is called ANOVA!
- You can think of “between-group” as marking variability explained by the predictor of interest and “within group” as marking everything else (residual).

# TABLE OF CONTENTS

- 1 ANOVA: Motivation
- 2 One-way ANOVA
- 3 Two-way ANOVA
- 4 Coefficient of determination

## Example: Factorial design

- Suppose we want to compare the effects of  $J$  chemotherapy agents together with  $K$  different dosage levels. In total, there are  $J \times K$  combinations of levels. A total of  $n_{jk}$  patients are assigned to each group, with  $N = \sum_{1 \leq j \leq J; 1 \leq k \leq K} n_{jk}$ .
- ANOVA model:  $Y_{ijk} = \mu + \alpha_{jk} + \epsilon_{ijk}$ .
  - ▶ What does the design matrix look like?
  - ▶ What is the interpretation of each of the following hypotheses?
    - ★  $H_0 : \alpha_{jk} = 0$  for all  $1 \leq j \leq J, 1 \leq k \leq K$ .
    - ★  $H_0 : \alpha_{11} = \alpha_{12} = \dots = \alpha_{1K}$ .
    - ★  $H_0 : \alpha_{11} = \alpha_{12} = \dots = \alpha_{1K}; \dots ; \alpha_{J1} = \alpha_{J2} = \dots = \alpha_{JK}$ .
    - ★  $H_0 : \alpha_{jk} = \alpha_j + \gamma_k$ .

# TWO-WAY ANOVA

## Example: Two-way ANOVA

- Suppose  $n_{jk} = n$ , so that  $N = n \times J \times K$ .
- A two-way ANOVA table typically would have the following information:

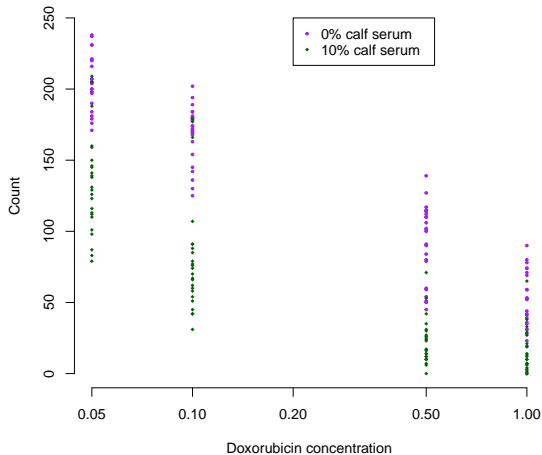
Source of variation	df	Sum of squares	Mean sum of squares
Factor X	$J - 1$	$SS_B^X$	$SS_B^X / (J - 1)$
Factor Z	$K - 1$	$SS_B^Z$	$SS_B^Z / (K - 1)$
Interaction XZ	$(J - 1)(K - 1)$	$SS_B^{XZ}$	$SS_B^{XZ} / ((J - 1)(K - 1))$
Error	$JK(n - 1)$	$SS_W$	$SS_W / (JK(n - 1))$
Total	$N - 1$	$SS_T$	$SS_T / (N - 1)$

## **Example:** Doxorubicin and serum

- Chemotherapy agents are generally preliminarily evaluated in a laboratory setting for activity.
- A sample drawn from a liquid culture of some cell line is exposed to a new drug or combinations of new drugs at varying concentrations.
- The cells are then put in a solid culture medium and incubated. The resulting colonies of cells can be counted using an optical scanner.
- The number of colonies visible at the end of incubation are a fair representation of the number of cells not killed by the treatment.
- As an example, consider a laboratory experiment of doxorubicin (concentrations of 0.05, 0.10, 0.50, and 1.00  $\mu\text{mol/L}$ ). Further, half the samples are exposed to 10% calf serum.
- Each of the eight defined groups has a sample size of  $n = 24$ .

# TWO-WAY ANOVA

## Example: Doxorubicin and serum



## Example: Doxorubicin and serum

```
zz <- aov(count ~ factor(doxconc) * factor(serum), data = dat)
```

```
> summary(zz)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
factor(doxconc)	3	558897	186299	285.870	< 2e-16	***
factor(serum)	1	214067	214067	328.480	< 2e-16	***
factor(doxconc):factor(serum)	3	16039	5346	8.204	3.74e-05	***
Residuals	184	119911	652			

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- How do we interpret each p-value?



## Concluding thoughts on ANOVA:

- There is a lot more that can be said about the ANOVA framework.
- The “balanced” case results in particularly nice properties.
- Special considerations are necessary for unbalanced cases.
- The assumption of equal variances is dreadful, and ANOVA doesn't provide a way out of it as easily as regression does.
- Mixed models/random effects models fall under the category of “ANCOVA” (**an**alysis of **cov**ariance). This falls under the umbrella of Advanced Regression Analysis II.
- I am biased toward the more flexible regression framework. However, ANOVA remains the approach of choice in some fields and you should familiarize yourself with the essential ideas.

# TABLE OF CONTENTS

- 1 ANOVA: Motivation
- 2 One-way ANOVA
- 3 Two-way ANOVA
- 4 Coefficient of determination

## **Decomposing variability:**

- Another topic that has been notably absent so far is the coefficient of determination (and the related correlation coefficient).
- It's easiest to motivate some of these measures of association once we've wrapped our minds around sources of variation.

# COEFFICIENT OF DETERMINATION

## Important decomposition:

- Consider the regression model  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ , with  $E[\boldsymbol{\epsilon}] = \mathbf{0}$  and  $\text{Cov}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{I}$ . Assume that  $\mathbf{X}$  is of full rank.
- Let  $\hat{y}$  denote the fitted value based on  $\hat{\boldsymbol{\beta}}$ , the OLS estimate of  $\boldsymbol{\beta}$ .
- Then,

$$\begin{aligned}SS_T &= \sum_{i=1}^N (y_i - \bar{y})^2 \stackrel{\text{math}}{=} \sum_{i=1}^N (y_i - \hat{y}_i)^2 + \sum_{i=1}^N (\hat{y}_i - \bar{y}_i)^2 \\ &= SS_R + SS_M.\end{aligned}$$

- $SS_T$ : total sum of squares.
  - ▶ Marks variability of  $Y$  about  $\bar{Y}$ .
- $SS_R$ : residual sum of squares (akin to “within”).
  - ▶ Marks variability of  $Y$  about  $\hat{Y}$ .
- $SS_M$ : model sum of squares (akin to “between”).
  - ▶ Marks variability of  $\hat{Y}$  about  $\bar{Y}$ .

## Additional decomposition:

- Note:  $SS_R = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y}$ . Why?
- With algebra,  $SS_R = \sum_{i=1}^N (y_i - \bar{y})^2 - \hat{\boldsymbol{\beta}}^T(\mathbf{X}^C)^T\mathbf{y}$ , where  $\mathbf{X}^C$  denotes the design matrix with predictors centered about their respective means.
- Taken together, this implies that

$$SS_M = \hat{\boldsymbol{\beta}}^T(\mathbf{X}^C)^T\mathbf{y}.$$

- Now, since  $\hat{\boldsymbol{\beta}}$  solves the appropriate normal equations, we further find that  $SS_M = \hat{\boldsymbol{\beta}}^T[(\mathbf{X}^C)^T\mathbf{X}^C]\hat{\boldsymbol{\beta}}$ .

# COEFFICIENT OF DETERMINATION

## Interpretation:

- The coefficient of determination is given as:

$$R^2 = \frac{SS_T - SS_R}{SS_T} = \frac{SS_M}{SS_T} = \frac{\widehat{\boldsymbol{\beta}}^T [(\mathbf{X}^C)^T \mathbf{X}^C] \widehat{\boldsymbol{\beta}}}{\sum_{i=1}^N (y_i - \bar{y})^2}$$

- Marks the proportion of “total variation” in the outcome explained by the overall model.
- Sometimes described as a measure of model fit, although I find this ambiguous. I want to avoid conflating concepts of correct model specification and a model’s predictive capacity.
- Now,  $SS_T$  can also be partitioned as:

$$\begin{aligned} SS_T &= \sum_{i=1}^N (y_i - \bar{y})^2 = \mathbf{y}^T \mathbf{y} - N\bar{y}^2 = (\mathbf{y}^T \mathbf{y} - \widehat{\boldsymbol{\beta}} \mathbf{X}^T \mathbf{y}) + (\widehat{\boldsymbol{\beta}} \mathbf{X}^T \mathbf{y} - N\bar{y}^2) \\ &= SS_R + SS_M. \end{aligned}$$

# COEFFICIENT OF DETERMINATION

## Properties:

- 1  $0 \leq R^2 \leq 1$ .
- 2  $R = \sqrt{R^2}$  marks the correlation between  $Y$  and  $\hat{Y}$ .
- 3 Adding variables to a model cannot decrease the value of  $R^2$ .

## Distribution under the null:

- Consider the hypothesis  $H_0 : \beta_1 = \dots = \beta_{K-1} = 0$  (i.e., the predictors are not linearly associated with the outcome).
- The  $F$ -test associated with this hypothesis can be written as:

$$F = \frac{R^2/(K-1)}{(1-R^2)/(N-K)}$$

- With a little algebra, we can rearrange, finding:

$$R^2 = \frac{(K-1)F}{(N-K) + (K-1)F} = \frac{\left(\frac{K-1}{N-K}\right)F}{1 + \left(\frac{K-1}{N-K}\right)F}$$



# COEFFICIENT OF DETERMINATION

**Distribution under the null:**  $H_0 : \beta_1 = \cdots = \beta_{K-1} = 0$

- If we are willing to assume that  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ , then we know that  $F \sim F_{K-1, N-K}$  under  $H_0$ , from which we learn that:

$$R^2 = \frac{\left(\frac{K-1}{N-K}\right) F}{1 + \left(\frac{K-1}{N-K}\right) F} \sim \text{Beta}\left(\frac{K-1}{2}, \frac{N-K}{2}\right).$$

- From this, we also learn that (under  $H_0$ ):

$$E[R^2] = \frac{K-1}{N-1}.$$

- This is the motivation for the adjusted  $R^2$ , which will have expectation zero under the null.

## Adjusted $R^2$ :

- The adjusted  $R^2$  penalizes the number of parameters in the model:

$$R_{\text{adj}}^2 = 1 - (1 - R^2) \frac{N - 1}{N - K}$$

- In so doing, we have that (under  $H_0$ ),

$$E[R_{\text{adj}}^2] = 0$$

- Trade-off: adjusted  $R^2$  can take on a negative value.

## Coefficient of partial determination: Partial $R^2$

- Suppose we seek to understand the proportion of variation not explained in a reduced model that is explained by predictors in a full(er) model.
- The partial  $R^2$  is given by:

$$R^2_{\text{Full|Reduced}} = \frac{SS_R^{\text{Reduced}} - SS_R^{\text{Full}}}{SS_R^{\text{Reduced}}}.$$

- Here,  $SS_R^{\text{Reduced}}$  is taking the place of  $SS_T$ , which you can think of as being based on a model that has been reduced to the maximal extent.

## Relationship to more general $F$ -test:

- Similarly, suppose we seek to test (intercept *not* involved):

$$H_0 : \begin{bmatrix} 0 & \mathbf{C}_{Q \times (K-1)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta}_{-0} \end{bmatrix} = \mathbf{0}$$

- Then, the  $F$ -statistic can be expressed as:

$$F = \frac{(R^2 - R_H^2)/Q}{(1 - R^2)/(N - K)},$$

where  $R_H^2$  is the coefficient of determination under the restricted model. You should be able to verify that this is the more general version of what we just saw.

- Keep in mind: assuming  $\text{rank}(\mathbf{C}) = Q$  and  $\mathbf{X}$  is of full rank.

**So far:**

- ANOVA.

## Up next:

- Various forms of model misspecification.