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: 10/03/2023

Table of Contents

ANOVA: Motivation

Ideas:

- We have talked about ANOVA previously (generally as a source of rank-deficient design matrices).
- ANOVA is an abbreviation for **an**alysis **o**f **va**riance. 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.

ANOVA: Motivation

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: $i = 1, \ldots, n_j$ observations subgroups $j = 0, 1$.
  - Possible model: $Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$.
  - Possible model: $Y_{ij} = \alpha_j + \epsilon_{ij}$.
- Example: $i = 1, \ldots, n_{kj}$ observations in subgroups defined by $j = 1, \ldots, J$ and $k = 1, \ldots, K$.
  - Possible model: $Y_{ijk} = \mu_{jk} + \epsilon_{ijk}$.
  - Possible model: $Y_{ijk} = \alpha_j + \beta_k + \epsilon_{ijk}$.

Table of Contents

One-way ANOVA

Example: One-way ANOVA with two groups

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

	Observations	Mean
Population 1	$y_{11}, \ldots, y_{1n_1}$	$\overline{y}_{1\cdot}$
Population 2	$y_{21}, \ldots, y_{2n_2}$	$\overline{y}_{2\cdot}$
Overall		$\overline{y}_{\cdot\cdot}$

- 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:

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

Example: One-way ANOVA with J groups

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

	Observations	Mean
Population 1	$y_{11}, \ldots, y_{1n_1}$	$\overline{y}_{1\cdot}$
Population 2	$y_{21}, \ldots, y_{2n_2}$	$\overline{y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$
Population J	$y_{J1}, \ldots, y_{Jn_J}$	$\overline{y}_{J\cdot}$
Overall		$\overline{y}_{\cdot\cdot}$

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

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

One-way ANOVA

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.

One-way ANOVA

Example: One-way ANOVA with J groups

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

$$\begin{aligned} \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ji} - \overline{y}_{..})^2 &= \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ji} - \overline{y}_{j.})^2 + \sum_{j=1}^{J} n_j (\overline{y}_{j.} - \overline{y}_{..})^2 \\ SS_T &= SS_W + SS_B \end{aligned}$$

- 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.

One-way ANOVA

Example: One-way ANOVA with J groups

- Suppose $n_1 = \cdots = n_J =: n$, 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

Two-way ANOVA

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 $1 \leq j \leq J$, $1 \leq k \leq K$.
    - $H_0 : \alpha_{11} = \alpha_{12} = \cdots = \alpha_{1K}$.
    - $H_0 : \alpha_{11} = \alpha_{12} = \cdots = \alpha_{1K}; \cdots ; \alpha_{J1} = \alpha_{J2} = \cdots = \alpha_{JK}$.
    - $H_0 : \alpha_{jk} = \alpha_j + \gamma_k$.

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:

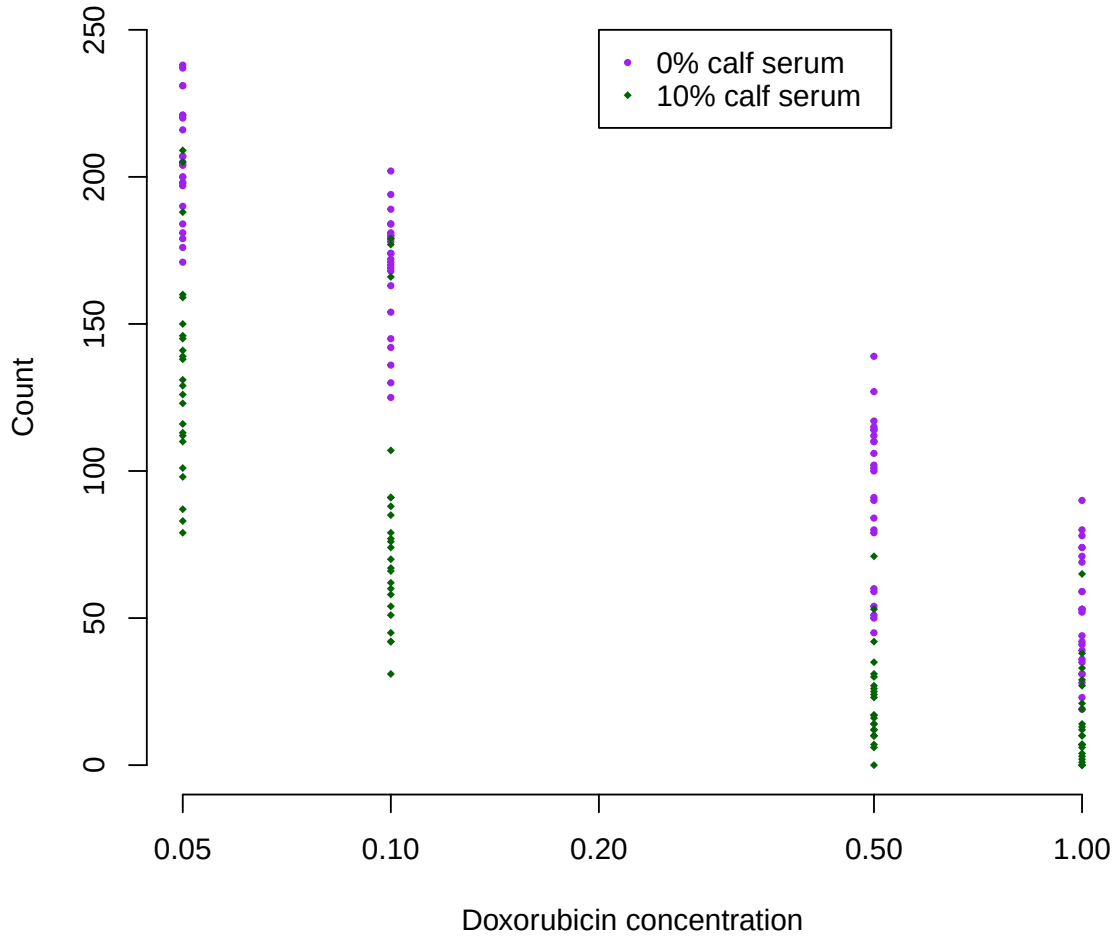
Source of variation	df	Sum of squares	Mean sum of squares
Factor X	$J-1$	SS_B^X	$SS_B^X/(J-1)$
Factor Z	$K-1$	SS_B^Z	$SS_B^Z/(K-1)$
Interaction XZ	$(J-1)(K-1)$	SS_B^{XZ}	$SS_B^{XZ}/((J-1)(K-1))$
Error	$JK(n-1)$	SS_W	$SS_W/(JK(n-1))$
Total	$N-1$	SS_T	$SS_T/(N-1)$

Two-way ANOVA

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 μmol/L). Further, half the samples are exposed to 10% calf serum.
- Each of the eight defined groups has a sample size of $n = 24$.

Example: Doxorubicin and serum



Two-way ANOVA

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**anlysis of **cova**riance). 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

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

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

- SS_T: total sum of squares.
  - Marks variability of Y about $\overline{Y}$.
- SS_R: residual sum of squares (akin to "within").
  - Marks variability of Y about $\widehat{Y}$.
- SS_M: model sum of squares (akin to "between").
  - Marks variability of $\widehat{Y}$ about $\overline{Y}$.

Coefficient of determination

Additional decomposition:

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

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

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

Coefficient of determination

Interpretation:

- The coefficient of determination is given as:

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

- 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:

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

Coefficient of determination

Properties:

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

Coefficient of determination

Distribution under the null:

- Consider the hypothesis $H_0 : \beta_1 = \cdots = \beta_{K-1} = 0$ (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):

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

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

Coefficient of determination

Adjusted R^2:

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

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

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

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

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

Coefficient of determination

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:

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

- Here, SS_R^{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.

Coefficient of determination

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.

Summary

So far:

- ANOVA.

Summary

Up next:

- Various forms of model misspecification.