CSCI 3022: Intro to Data Science

Lecture 24: ANOVA

Rachel Cox

Department of Computer Science

Announcements & Reminders

- Practicum 2 is posted. Due date - Friday May 1 at 11:59 pm
- Final Exam - May 5
  - 24 hour window to take exam.
  - More info to follow on Piazza

What will we learn today?

- ❑ **An**alysis **o**f **Va**riance ANOVA
- ❑ Formally test for a difference among lots of group means

- ❑ *Introduction to Statistical Learning, Chapter 3*

Review

We now have an idea of how to make hypothesis tests on our MLR model. Each MLR comes with a lot of p-values:

- An F statistic tells us whether our model as a whole is significantly useful.
- A T statistic for each and every β testing whether its nonzero in the presence of the other linear terms.

The Partial F

The F distribution allows us to compare variances of different models.

Full model: $Y = \beta_0 + \beta_1 X_1 + \ \ldots + \beta_4 X_4$ •
Reduced model: $Y = \beta_0 + \beta_2 X_2 + \beta_4 X_4$ •

Null: H_0: $\beta_1 = \beta_3 = 0$

Alternative: Either/both of $\beta_1 \neq 0$ or $\beta_3 \neq 0$. Alternatively, the overall model captures significantly more variability in Y by including both β_1 and β_3.

- In many situations, one first builds a model containing p predictors and then wishes to know whether any of the predictors in a particular subset provide useful information about Y.

ANOVA

The F distribution is commonly used for a type of analysis that overlaps with MLR: ANOVAs of **An**alysis **o**f **Va**riance.

An ANOVA typically refers to performing inference on whether categorical features in the data are important. These may include:

- Binary outcomes
- Categorical outcomes
- Artificial stratifications of the data

ANOVA

We're often interested in comparing the means from different groups.

Example: Suppose we're tasked with a weight loss study. In this study, we have three groups:

Control group: exercise only
Treatment A: exercise plus Diet A
Treatment B: exercise plus Diet B

We record the weight-loss of each participant after one week of the study and find the following:

Control	Diet A	Diet B
3	5	5
2	3	6
1	4	7

ANOVA

We're often interested in comparing the means from different groups.

Example: (continued) Are the means if the different groups all the same? What would we do if there were only two groups.

Control	Diet A	Diet B
3	5	5
2	3	6
1	4	7

- A linear model with only categorical predictors have been traditionally called analysis of variance (ANOVA).

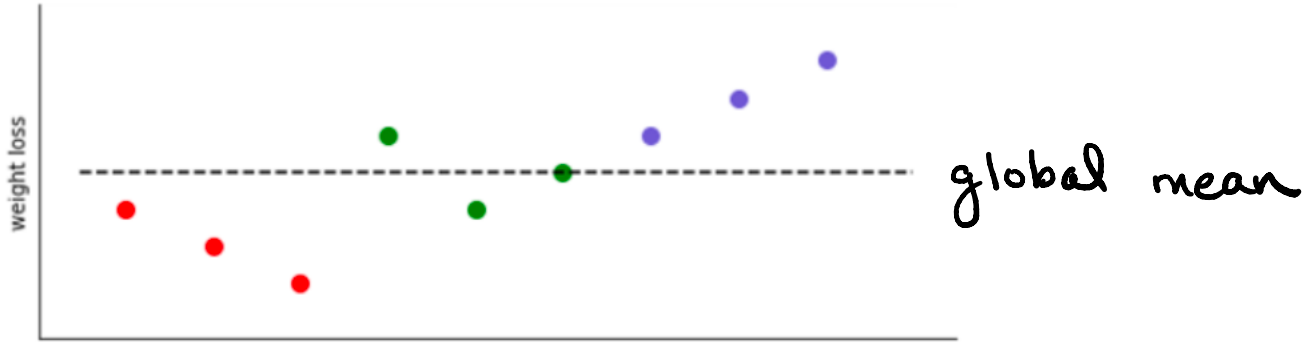
- The purpose of ANOVA is to determine whether there are any statistically significant differences between the means of several independent groups.

ANOVA

The one-way ANOVA model can be used to test the null hypothesis:
H_0: $\mu_A = \mu_B = \mu_C =$ ... for all groups/categories

As a result of the null hypothesis being "all is equal", we will compare what happens in a model with different means to a model with one global mean.

So we will find a global mean and the individual group means.



control group - red
Diet A - green
Diet B - blue

ANOVA

The one-way ANOVA model can be used to test the null hypothesis:
$H_0 = \mu_A = \mu_B = \mu_C =$ … for all groups/categories

Global mean = $\bar{\bar{y}} = \frac{1}{\#\text{ data points}} \cdot \sum (\text{all data points})$

$= \frac{1}{9}(3+2+1+5+3+4+5+6+7)$

$= \frac{36}{9} = 4$

$\bar{\bar{y}} = 4$

Control	Diet A	Diet B
3	5	5
2	3	6
1	4	7

Group means =

$\bar{y}_1 = \frac{1}{3}(3+2+1) = 2$

$\bar{y}_2 = \frac{1}{3}(5+3+4) = \frac{12}{3} = 4$

$\bar{y}_3 = \frac{1}{3}(5+6+7) = \frac{18}{3} = 6$

$\bar{y}_1 = 2 \quad \bar{y}_2 = 4 \quad \bar{y}_3 = 6$

ANOVA

After this, we look at where the variance in the data is. Specifically, we want to compute the sum of squares… but we want to split it up. Suppose we have I groups each with n_i points (maybe unequal).

Total sum of squares: $SST = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{\bar{y}}\right)^2$

Within-group sum of squares, measuring how much groups are split from their own mean:

$$SSW = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2$$

★ **Between-group** sum of squares, measuring how much groups are split from the global mean:

$$SSB = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{\bar{y}})^2$$

global mean

each groups mean

ANOVA

Where does the total variation in the data come from?

Look first at the total sum of squares: $SST = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{\bar{y}}\right)^2$

A helpful decomposition: $y_{ij} - \bar{\bar{y}} = \left(y_{ij} - \bar{y}_i\right) + \left(\bar{y}_i - \bar{\bar{y}}\right)$

$$SST = SSW + SSB$$

ANOVA

Let's compute the variances for our data!

$\bar{\bar{y}} = 4$

$\bar{y}_1 = 2$

$\bar{y}_2 = 4$

$\bar{y}_3 = 6$

Control	Diet A	Diet B
3	5	5
2	3	6
1	4	7

The between groups sum of squares is:

$$SSB = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\bar{y}_{ij} - \bar{\bar{y}})^2$$

$$= (\bar{y}_{11} - \bar{\bar{y}})^2 + (\bar{y}_{12} - \bar{\bar{y}})^2 + (\bar{y}_{13} - \bar{\bar{y}})^2 +$$

$$+ (\bar{y}_{21} - \bar{\bar{y}})^2 + (\bar{y}_{22} - \bar{\bar{y}})^2 + (\bar{y}_{23} - \bar{\bar{y}})^2 +$$

$$+ (\bar{y}_{31} - \bar{\bar{y}})^2 + (\bar{y}_{32} - \bar{\bar{y}})^2 + (\bar{y}_{33} - \bar{\bar{y}})^2 +$$

$$= 3(\bar{y}_1 - \bar{\bar{y}})^2 + 3(\bar{y}_2 - \bar{\bar{y}})^2 + 3(\bar{y}_3 - \bar{\bar{y}})^2$$

$$= 3(2-4)^2 + 3(4-4)^2 + 3(6-4)^2$$

$$= 3 \cdot 4 + 3 \cdot 0 + 3 \cdot 4$$

$$\boxed{SSB = 24}$$

ANOVA

Let's compute the variances for our data!

$\bar{y}_1 = 2$
$\bar{y}_2 = 4$
$\bar{y}_3 = 6$

$$SSW = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

$$= (y_{11} - \bar{y}_1)^2 + (y_{12} - \bar{y}_1)^2 + (y_{13} - \bar{y}_1)^2$$

$$+ (y_{21} - \bar{y}_2)^2 + (y_{22} - \bar{y}_2)^2$$

$$+ (y_{23} - \bar{y}_2)^2$$

The within groups sum of squares is:

Control	Diet A	Diet B
3	5	5
2	3	6
1	4	7

$$+ (y_{31} - \bar{y}_3)^2 + (y_{32} - \bar{y}_3)^2 + (y_{33} - \bar{y}_3)^2$$

$$= (3-2)^2 + (2-2)^2 + (1-2)^2 + (5-4)^2 + (3-4)^2$$
$$+ (4-4)^2 + (5-6)^2 + (6-6)^2 + (7-6)^2$$

$$= 1 + 0 + 1 + 1 + 1 + 0 + 1 + 0 + 1$$

$$= 6$$

$SSW = 6$

ANOVA

$\bar{\bar{y}} = 4$

Let's compute the variances for our data!

$$SST = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}})^2$$

The between groups sum of squares is:

$$= (3-4)^2 + (2-4)^2 + (1-4)^2 + (5-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (6-4)^2 + (7-4)^2$$

Control	Diet A	Diet B
3	5	5
2	3	6
1	4	7

The within groups sum of squares is:

$$= 1 + 4 + 9 + 1 + 1 + 0 + 1 + 4 + 9$$

$$= 30$$

The total sum of squares is:

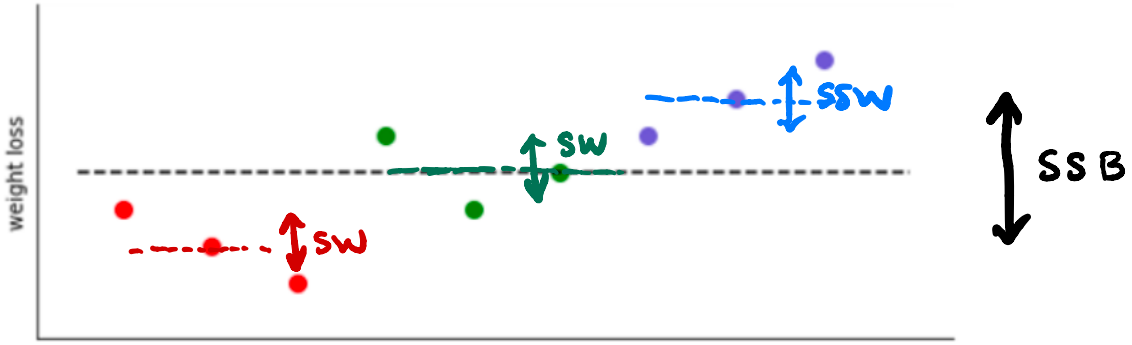
$SST = 30$

$SSB = 24$ $\quad$ $SSW = 6$

ANOVA

So we have $SSB = 24$, $SSW = 6$, and $SST = 30$

This means there is 30 units total of summed squared "movement" of the data. 24 of that is attributed to the 3 groups each having different means. 6 of that is attributed to movement of the random variable within each group. This means that allowing each group to have it's own mean accounts for $\frac{SSB}{SST} = \frac{24}{30} = 80\%$ of the variability. That's an R^2.



Control	Diet A	Diet B
3	5	5
2	3	6
1	4	7

ANOVA

What about **degrees of freedom**?

	Control	Diet A	Diet B
0	3	5	5
1	2	3	6
2	1	4	7

- The **between groups** degrees of freedom is: $SSB_{df} = I - 1$

$$SSB = \sum_{i=1}^{I} n_i(\overline{y}_i - \overline{\overline{y}})^2$$

I : # of groups.

- The **within groups** degrees of freedom is: $SSW_{df} = N - I$

$$SSW = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2$$

N = total number of data points across all groups

$SSW_{df} = I(n_i - 1)$

# of groups; degrees of freedom for each group

- So for our example, we have: $SSB_{df} = 3-1 = 2$ and $SSW_{df} = 9 - 3 = 6$

ANOVA

We want to perform a hypothesis test to determine if the group means are equal. We have…

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

H_1: $\mu_i \neq \mu_j$ for some pair i, j

Our test statistic will be: $F = \dfrac{SSB / SSB_{df}}{SSW / SSW_{df}} \sim F_{I-1,N-I}$

$= \dfrac{\frac{SSB}{I-1}}{\frac{SSW}{N-I}}$

Rejection region:

$F \geq F_{\alpha, I-1, N-I}$ = stats.f.ppf(1-α, dfn=I-1, dfd=N-I)

p-value:

1-stats.f.cdf(F, I-1, N-I)

ANOVA

In reporting results, it's common practice to stick all of these squared terms and degrees of freedom into a table.

	Control	Diet A	Diet B
0	3	5	5
1	2	3	6
2	1	4	7

sum of squares

degrees of freedom

ANOVA	SS	DF	SS/DF	F_{stat}
between	24	2	12	12/1 = 12
within	6	6	1	
total	30	8		

N - 1

Suppose $\alpha = .05$

Exercise for home

Next thing: F_{crit} = stats.f.ppf(.95, 2, 6)

ANOVA as a multiple linear regression

We can write our end result that group matter into a linear model.

e.g. $y =$ Control + effects of group + errors

A point in group 1 would look like:

$$y = \beta_0 + \beta_1 + \epsilon$$

A data point in group 2 would look like:

$$y = \beta_0 + \beta_2 + \epsilon$$

We use indicator variables to set this up:

$$x_{ij} = \begin{cases} 1 & \text{if data point } j \text{ is in group } i \\ 0 & \text{otherwise} \end{cases}$$

ANOVA as a multiple linear regression

- Choose one group as the control.
- Model: $y_{ij} = \mu_0 + \tau_1 x_{1j} + \tau_2 x_{2j} + \ \ldots + \tau_{I-1} x_{I-1,j} + \epsilon_{ij}$

y_{ij} is the j^{th} response for the i^{th} group

$$x_{ij} = \begin{cases} 1 & \text{if data point } j \text{ is in group } i \\ 0 & \text{otherwise} \end{cases}$$

Control	Diet A	Diet B
3	5	5
2	3	6
1	4	7

y_{ij}	x_{1j}	x_{2j}
3	0	0
2	0	0
1	0	0
5	1	0
3	1	0
4	1	0
5	0	1
6	0	1
7	0	1

ANOVA as a multiple linear regression

- Model: $y_{ij} = \mu_0 + \tau_1 x_{1j} + \tau_2 x_{2j}$

Control	Diet A	Diet B
3	5	5
2	3	6
1	4	7

Mean response for control:

$$x_{1j} = x_{2j} = 0 \rightarrow y_{ij} = \mu_0$$

Mean response for Diet A:

$$x_{2j} = 0 \rightarrow y_{ij} = \mu_0 + \tau_1$$

Mean response for Diet B:

$$x_{1j} = 0 \rightarrow y_{ij} = \mu_0 + \tau_2$$

ANOVA as a multiple linear regression

The F-test gives us a single conclusion on a model like this: groups matter. It doesn't tell us how much or which groups matter. To answer this, we have to resort to more individual tests, like testing group A against group B, group A against the control, and group B against the control.

Suppose we perform these 3 tests. Each is a t test with error probability $\alpha = .05$. What is the probability we commit an error?

- $P(no\ error) = P(no\ error\ Test\ \#1) \cdot P(no\ error\ Test\ \#2) \cdot P(no\ error\ Test\ \#3) = (.95)^3$

$P(error) = 1 - .95^3 \approx .14$

ANOVA as a multiple linear regression

- Use α for the F-statistic that tells us groups matter.

- If and only if the F-statistic is significant, perform a special, very careful version of pairwise testing between our groups that accounts for both the fact that we're doing multiple tests and that those tests aren't independent.

Tukey's honest significance test

Suppose we determine that some of the means are different. How can we tell which ones?

Tukey's Honest Significance Test (HST) - aka Tukey's Range Test aka Tukey's Honest Significant Difference (HSD)

- Hypothesis test for pairwise comparison of means
  - It's many pairwise tests using what's called the studentized range distribution

- Adjusts so that the probability of making a Type I error over all possible pairwise comparisons = α

Next Time:

- Logistic Regression