CSCI 3022-002 Intro to Data Science
ANOVA

Announcements and Reminders

- Homework due tonight!
- Kinda-sorta-optional notebook day Wednesday.

Where we at?

At this point, our MLR workflow looks like:

1. Plot the linear model
2. See if some predictors are redundant
3. Plot residuals of linear model, check for **normality**, **independence**, **structure**.
4. Hit model with a math-shaped stick to fix these problems.

We've also got some idea on how to make hypothesis tests on our model. Each MLR comes with a lot of p-values:

1. An F statistic telling us whether our model as a whole is significantly useful.
2. 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 also lent it self to more nuanced comparisons. Instead of just comparing the variance of our model to the variance of Y, we could compare variances of different models.

Null:

Alternative:

The Partial F

The F distribution also lent it self to more nuanced comparisons. Instead of just comparing the variance of our model to the variance of Y, we could compare variances of different models.

Full model: $\underline{Y} = \beta_0 + \beta_1 \underline{X_1} + \ldots \beta_4 \underline{X_4}$
Reduced Model: $\underline{Y} = \beta_0 + \beta_2 \underline{X_2} + \beta_4 \underline{X_4}$

Null:

Alternative:

The Partial F

Stat for comparing variances: $F : \frac{S_1^2}{S_2^2}$

The F distribution also lent it self to more nuanced comparisons. Instead of just comparing the variance of our model to the variance of Y, we could compare variances of different models.

Full model: $\underline{Y} = \beta_0 + \beta_1 \underline{X_1} + \ldots \beta_4 \underline{X_4}$ 4 features
Reduced Model: $\underline{Y} = \beta_0 + \beta_2 \underline{X_2} + \beta_4 \underline{X_4}$ vs only 2 of those

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.

The Partial F

The F distribution also lent it self to more nuanced comparisons. Instead of just comparing the variance of our model to the variance of Y, we could compare variances of different models.

Full model: $\underline{Y} = \beta_0 + \beta_1 \underline{X_1} + \ldots \beta_4 \underline{X_4}$
Reduced Model: $\underline{Y} = \beta_0 + \beta_2 \underline{X_2} + \beta_4 \underline{X_4}$

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.

The Partial F

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:

1. Binary outcomes T/F or Men/Women
2. Categorical outcomes Red/blue/green
3. Artificial stratifications of the data

people who make > $100,000/yr
" " " < $100,000/yr.

ANOVA

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

weight lost = f(treatment).

We find the following:

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

lbs lost since start of study

ANOVA

avg = 2 ↓ avg = 4 ↓ avg = 6 ↓

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

What can we conclude?

1. Why don't we just test Control vs. A then control vs. B then A vs. B as t-tests?
2. Are the means of each group the same?

↳ 3 tests is scary!

If $\alpha = .05$ per test

this is 3 chances to commit an error, $\approx 15\%$.

If we have 10 groups: $\binom{10}{2} = 45$ tests

One-Way ANOVA

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 several independent groups.

The one-way ANOVA model can be used to test the null hypothesis:

One-Way ANOVA

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 several independent groups.

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

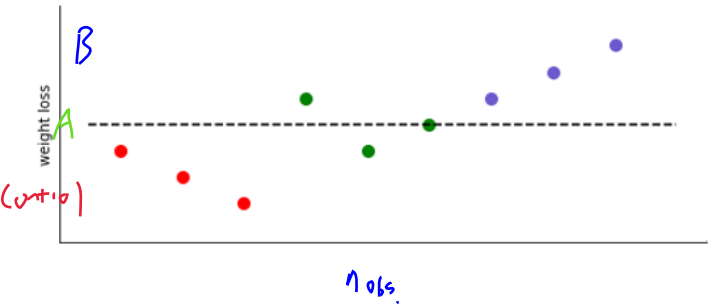
One-Way ANOVA

H_0: only one mean

H_A: allow different means between groups

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 find a global mean and some individual group means.



One-Way ANOVA

We find a global mean and some individual group means.

$$\bar{\bar{y}} = \frac{n_A \bar{y}_A + n_B \bar{y}_B + n_C \bar{y}_C}{n_A + n_B + n_C} = \frac{\sum \text{all } y}{n_y}$$

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

2 4 6

Global mean $\bar{\bar{y}}$: avg all (9) points = 4

Group means $[\bar{y}_1, \bar{y}_2, \bar{y}_3]$: 2 4 6

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!)

One-Way ANOVA

We find a global mean and some individual group means.

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

Global mean $\bar{\bar{y}}$:

Group means $[\bar{y_1}, \bar{y_2}, \bar{y_3}]$: [2,4,6]

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!)

One-Way ANOVA

We find a global mean and some individual group means.

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

Global mean $\bar{\bar{y}}$: 4

Group means $[\bar{y_1}, \bar{y_2}, \bar{y_3}]$: [2,4,6]

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!)

One-Way ANOVA

1. **Total** sum of squares:

$$\sum_{\text{all}} \left(\text{data} - \underline{\text{global}}\ \text{mean} \right)^2 = \sum (y_i - \bar{y})^2$$

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

$$\sum_{\text{all data (all groups)}} \quad \sum_{\text{within each group}} \left(\text{data} - \begin{matrix} \text{group} \\ \text{group} \\ \text{group} \end{matrix}\ \text{mean} \right)^2$$

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

$$\sum_{\text{all data}} \left(\text{(all/group means} - \text{global means)} \right)^2$$

One-Way ANOVA

1. **Total** sum of squares:

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

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

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

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

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

Some squares

$$R^2 = 1 - \frac{\text{error}^2}{\text{total}^2}$$

Maybe intuitively, SST=SSB+SSW. This results from rewriting the SST inside part with

$$y_{ij} - \bar{\bar{y}} = \underbrace{y_{ij} - \bar{y}_i}_{within} + \underbrace{\bar{y}_i - \bar{\bar{y}}}_{between}$$

(error: variability not captured by group itself)

This lets us perform the same comparisons of variance that went into things like R^2 and the F test from MLR. Now, SSW represents some kind of error term (the variance that we *can't* capture with localized means) and SSB represents the amount that our model deviates from a baseline model if we do use localized means: the same as SSR from MLR!

One-Way ANOVA

We have $\bar{\bar{y}} = 4$ and $[\bar{y_1}, \bar{y_2}, \bar{y_3}] = [2, 4, 6]$

	Control	Diet A	Diet B
0	3	5	5
1	2	3	6
2	1	4	7
$\bar{y}_{group}$	2	4	6

SST = 30: Idea: Y "moves" 30 (squared) units total

24 of these from group means, 6 are from points within groups.

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

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

$3(2-4)^2 + 3(4-4)^2 + 3(6-4)^2 = 24$

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

$(3-2)^2 + (2-2)^2 + (1-2)^2$

$(5-4)^2 + (4-4)^2 + (4-4)^2$

$(5-6)^2 + (6-6)^2 + (7-6)^2$

$= 6.$

One-Way ANOVA

We have $\bar{\bar{y}} = 4$ and $[\bar{y_1}, \bar{y_2}, \bar{y_3}] = [2, 4, 6]$

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

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

$$= 3(2-4)^2 + 3(4-4)^2 + 3(6-4)^2 = 24$$

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

One-Way ANOVA

We have $\bar{\bar{y}} = 4$ and $[\bar{y_1}, \bar{y_2}, \bar{y_3}] = [2, 4, 6]$

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

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

$$= 3(2-4)^2 + 3(4-4)^2 + 3(6-4)^2 = 24$$

SSW$= \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y_i})^2 = \left[(3-2)^2 + (2-2)^2 + (1-2)^2\right]$

$$+ \left[(5-4)^2 + (3-4)^2 + (4-4)^2\right] + \left[(5-6)^2 + (6-6)^2 + (7-6)^2\right] = 6$$

One-Way ANOVA

We have $\bar{\bar{y}} = 4$ and $[\bar{y_1}, \bar{y_2}, \bar{y_3}] = [2, 4, 6]$

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

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

$$= 3(2-4)^2 + 3(4-4)^2 + 3(6-4)^2 = 24$$

SSW$= \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y_i})^2 = \left[(3-2)^2 + (2-2)^2 + (1-2)^2\right]$

$$+ \left[(5-4)^2 + (3-4)^2 + (4-4)^2\right] + \left[(5-6)^2 + (6-6)^2 + (7-6)^2\right] = 6$$

One-Way ANOVA

We have $SSB = 24; SSW = 6$

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

As a result, $SST = 30$. 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 $SSB/SST = 24/30 = 80\%$ of the variability. That's an R^2.

One-Way ANOVA

If it's an R^2, it's also a test!

Null hypothesis:

Alternative hypothesis:

One-Way ANOVA

SST: Sum of Squares total (variability of all data)

SSB: Sum Squares between groups (how different are the groups)

SSR: Sum Squares residuals (error or unaccounted randomness after using groups.

If it's an R^2, it's also a test!

Null hypothesis:

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

Alternative hypothesis:

$$H_a : \mu_i \neq \mu_j \quad \text{for at least one pair } i, j.$$

not multiple tests

One-Way ANOVA

If it's an R^2, it's also a test!

Null hypothesis:

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

Alternative hypothesis:

$$H_a : \mu_i \neq \mu_j \qquad \text{for at least one pair } i, j.$$

Test statistic value:

$$F = \frac{SSB/SSB_{dof}}{SSW/SSWdof} = \frac{SSB/(I-1)}{SSW/(N-I)} \sim F_{I-1,N-I}$$

Rejection region for a level test: $f \geq F_{\alpha,I-1,N-I}$

What are our assumptions?

N : total data

I : # of groups.

One-Way ANOVA

If it's an R^2, it's also a test!

Null hypothesis:

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

Alternative hypothesis:

$$H_a : \mu_i \neq \mu_j \qquad \text{for at least one pair } i, j.$$

Test statistic value:

$$F = \frac{SSB/SSB_{dof}}{SSW/SSWdof} = \frac{SSB/(I-1)}{SSW/(N-I)} \sim F_{I-1,N-I}$$

Rejection region for a level test: $f \geq F_{\alpha,I-1,N-I}$

What are our assumptions?

Independence between data values both within and across groups, **normality** of variances in both SSB and SSW (looks identical with-group variances)

ANOVA tables

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

Sums of squares

Degrees freedom

ANOVA	SS	DF	SS/DF	F_{stat}
between				
within				
total				

compare 2 variances

a variance

ANOVA tables

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

9 data points
3 groups

ANOVA	SS	DF	SS/DF	F_{stat}
between	24	3-1=2		
within	6	9-3=6		
total	30	8		

ANOVA tables

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

F: SS/DF of between / SS/DF of within

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

ANOVA	SS	DF	SS/DF	F_{stat}
between	24	3-1=2	12	12
within	6	9-3=6	1	$p(12) = .008$
total	30	8		

.008 < α ? then reject!

Déjà vu

Did this all look familiar? Here's the MLR version of the same test. Null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

Alternative hypothesis:

$$H_a : \text{at least one } \beta_j \neq 0.$$

Test statistic value:

$$F = \frac{SSR/(p+1)}{SST/(n-p+1)}$$

Rejection region for a level test: $f \geq F_{\alpha,p+1,n-(p+1)}$

Linear ANOVAs

We can write our end result: **groups matter!** into a linear model. It might look something like: $y =$Control $+$ effects of group $+$errors

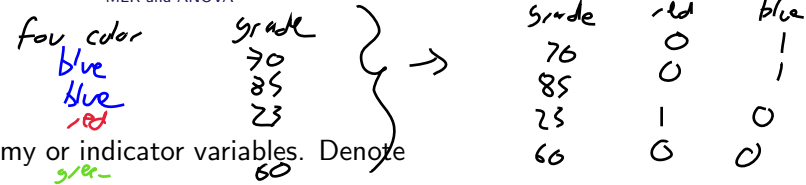
So a data point in group 1 would look like:

$$y = \beta_0 + \beta_1 + \varepsilon = \beta_0 + \beta_1 + \varepsilon$$

and a data point in group 2 would look like:

$$y = \beta_0 + \beta_2 + \varepsilon = \beta_0 + \beta_2 + \varepsilon$$

Linear ANOVAs



More formally, we use dummy or indicator variables. Denote

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

1=TRUE
0=FALSE

If we use this, we can answer the more general case. We choose one group as a control group, and the entire model becomes

$$y_{ij} = \underbrace{\mu_0}_{\text{control mean}} + \beta_1 x_{1,j} + \underbrace{\beta_2}_{\text{grp 2 offset}} x_{2,j} + \cdots + \tau_{I-1} \underbrace{x_{I-1,j}}_{\text{0 outside group I-1}} + \varepsilon$$

Linear ANOVAs

The matrix form of x may be more appealing:

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

y_{ij}	x_{Aj}	x_{Bj}

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

Linear ANOVAs

The matrix form of x may be more appealing:

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

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

y_{ij}	x_{Aj}	x_{Bj}	
3	0	0	Control
2	0	0	
1	0	0	
5	1	0	A
3	1	0	
4	1	0	
5	0	1	B
6	0	1	
7	0	1	

X_{height}
6'
5'10
4'8.2

Linear ANOVA Means

The means of each group now occur when we ask what happens to the data points in those groups. Notation: $\mathbb{1}_E$ is used as an indicator or dummy variable. It equals 1 when the event E is true, else 0. "truth" variable / indicator / dummy

$$y_i = \beta_0 + \beta_1 \mathbb{1}_{x_i \in A} + \beta_2 \mathbb{1}_{x_i \in B} + \varepsilon$$

baseline

$y_i = \beta_0 + \beta_1 \mathbb{1}_{x_i \in F}$

β_0: mean for F is false

$\beta_0 + \beta_1$: mean for F is true

For x in the control group, $E[Y_i | x_i \in C] = \beta_0 + 0 + 0$.
For x in group A, $E[Y_i | x_i \in A] = \beta_0 + \beta_1 + 0$.
For x in group B, $E[Y_i | x_i \in B] = \beta_0 + 0 + \beta_2$. We call β_1 and β_2 **treatment effects.**

Multiple testing

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 do this, we have to resort to more individual tests, like testing A against B, A against control, B against 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?

Multiple testing

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 do this, we have to resort to more individual tests, like testing A against B, A against control, B against 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?

An error is the union of (Error on Test # 1), (Error on Test # 2), and (Error on Test # 3), each of which occur 5% of the time. This is a good opportunity for Demorgan's Laws!

Multiple testing

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 do this, we have to resort to more individual tests, like testing A against B, A against control, B against 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?

$P(\text{no error}) = P(\text{no error on Test } \#1)P(\text{no error on Test } \#2)P(\text{no error on Test } \#3) = (.95)^3$ so $P(\text{error}) = 1 - .95^3$

Multiple testing

A quick heuristic approximation is that if we're performing k tests and α is small, we increase our *true* error rate by $k\alpha$. This means that dividing our α by k would "fix" this issue... but that could make the rejection region very small. So we:

1. Use α for the F-statistic that tells us **groups** matter.
2. 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.

Multiple testing

A quick heuristic approximation is that if we're performing k tests and α is small, we increase our *true* error rate by $k\alpha$. This means that dividing our α by k would "fix" this issue... but that could make the rejection region very small. So we:

1. Use α for the F-statistic that tells us **groups** matter.
2. 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.
   Why not independent?? If $A \neq B$ and $C = A$, do we *really* have to test $B = C$?

Multiple testing

A quick heuristic approximation is that if we're performing k tests and α is small, we increase our *true* error rate by $k\alpha$. This means that dividing our α by k would "fix" this issue... but that could make the rejection region very small. So we:

1. Use α for the F-statistic that tells us **groups** matter.
2. 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.
   Why not independent?? If $A \neq B$ and $C = A$, do we *really* have to test $B = C$?
   Sike! We do. If C is somewhere in-between A and B we could find that we can statically differentiate A and B but not C from either. But that test isn't full independent, because we're less likely to find $B = C$ given our existing information.

Multiple testing

The special test for performing **post hocs** (after significance analysis) is called a **Tukey** test, or Tukey HSD. The Tukey HSD ("honest significant difference") tells us which groups are different from other groups, and can lead to a table of pairwise comparisons.

In statistical reporting, once we're sure that our parameters/groupings matter, we can "finish" our problem by showing plots of the means/boxplots of the different groups. If two groups aren't demonstrably different, our final results should group them together before creating final boxplots. Then we can be honest, too!