

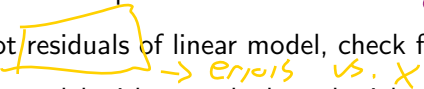

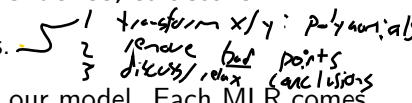



CSCI 3022 Intro to Data Science

ANOVA and MLR Wrapup

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. 

Announcements and Reminders

- ▶ Final weeks' schedule: Today: ANOVA. Wednesday: (Logistic Regression) ^{untested} brief overview of notebook content. Reading Day: **optional** Review session, no class. Late: Zach will publish an *untested/optional* lecture on stochastic gradient optimization.
- ▶ Deadlines: Pen-and-paper exam and Practicum: May 2 (sorry for the delayed posting on pen-and-paper exam! Hopefully you used that time ~~in~~ to work on the practicum!)

See Canvas/schedule for

1) Exam

2) Practicum

3) past Exams (modules).

Test for Equivalence of Variance

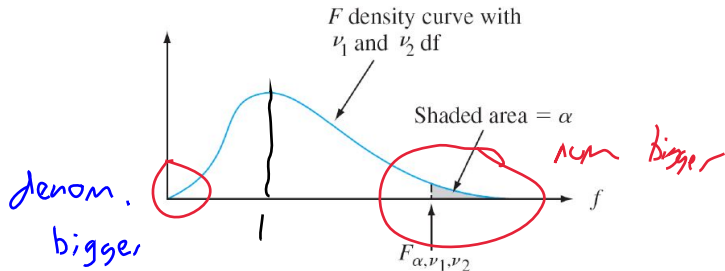
Variance: #1: σ_1^2 vs. #2: σ_2^2

The F probability distribution has two parameters, denoted by ν_1 and ν_2 . The parameter ν_1 is called the numerator degrees of freedom, and ν_2 is the denominator degrees of freedom.

~~$\sigma_1^2 - \sigma_2^2$~~ : instead σ_1^2 / σ_2^2 .

A random variable that has an F distribution cannot assume a negative value. The density function is complicated and will not be used explicitly, so it's not shown.

Figure below illustrates a typical F density function.:



Test for Equivalence of Variance

We use F_{α, ν_1, ν_2} for the value on the horizontal axis that captures of the area under the F density curve with ν_1 and ν_2 df in the upper tail.

The density curve is not symmetric, so it would seem that both upper- and lower-tail critical values must be tabulated. This is not necessary, though, because of the fact that

$$F_{1-\alpha, \nu_1, \nu_2} =$$

Test for Equivalence of Variance

We use F_{α, ν_1, ν_2} for the value on the horizontal axis that captures of the area under the F density curve with ν_1 and ν_2 df in the upper tail.

The density curve is not symmetric, so it would seem that both upper- and lower-tail critical values must be tabulated. This is not necessary, though, because of the fact that

$$F_{1-\alpha, \nu_1, \nu_2} = \frac{1}{F_{\alpha, \nu_1, \nu_2}}$$

For example, $F_{.05, 6, 10} = 3.22$ and $F_{.95, 10, 6} = 0.31 = 1/3.22$.

Test for Equivalence of Variance

A test procedure for hypotheses concerning the ratio σ_1^2/σ_2^2 is based on the following result.

Theorem:

Let X_1, X_2, \dots, X_m be a random sample from a normal distribution with variance σ_1^2 let Y_1, Y_2, \dots, Y_n be another random sample (independent of the X_i 's) from a normal distribution with variance σ_2^2 and let s_1^2 and s_2^2 denote the two sample variances. Then the rv

has an F distribution with $\nu_1 = m - 1$ and $\nu_2 = n - 1$. Because F involves a ratio rather than

a difference, the test statistic is the ratio of sample variances.

Test for Equivalence of Variance

A test procedure for hypotheses concerning the ratio σ_1^2/σ_2^2 is based on the following result.

Theorem:

Let X_1, X_2, \dots, X_m be a random sample from a normal distribution with variance σ_1^2 let Y_1, Y_2, \dots, Y_n be another random sample (independent of the X_i 's) from a normal distribution with variance σ_2^2 and let s_1^2 and s_2^2 denote the two sample variances. Then the rv

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

has an F distribution with $\nu_1 = m - 1$ and $\nu_2 = n - 1$. Because F involves a ratio rather than

a difference, the test statistic is the ratio of sample variances.

Test for Equivalence of Variance

Null hypothesis: H_0 :

Test statistic value:

<u>Alt Hypothesis</u>	<u>Rejection Region</u>
-----------------------	-------------------------

<u>p-value:</u>

Test for Equivalence of Variance

Null hypothesis: $H_0 : \sigma_1^2 = \sigma_2^2$

Test statistic value:

$$F = s_1^2 / s_2^2$$

Alt Hypothesis Rejection Region

p-value:

Test for Equivalence of Variance

Null hypothesis: $H_0 : \sigma_1^2 = \sigma_2^2$

Test statistic value:

$$F = s_1^2 / s_2^2$$

<u>Alt Hypothesis</u>	<u>Rejection Region</u>
$H_a : \sigma_1^2 > \sigma_2^2$	
$H_a : \sigma_1^2 < \sigma_2^2$	
$H_a : \sigma_1^2 \neq \sigma_2^2$	

p-value:

Test for Equivalence of Variance

Null hypothesis: $H_0 : \sigma_1^2 = \sigma_2^2$

Test statistic value:

$$F = s_1^2 / s_2^2$$

<u>Alt Hypothesis</u>	<u>Rejection Region</u>	<u>p-value:</u>
$H_a : \sigma_1^2 > \sigma_2^2$	$F_{stat} > F_{\alpha, m-1, n-1}$	
$H_a : \sigma_1^2 < \sigma_2^2$	$F_{stat} < F_{1-\alpha, m-1, n-1}$	
$H_a : \sigma_1^2 \neq \sigma_2^2$	$F_{stat} < F_{1-\alpha/2, m-1, n-1}$ OR $F_{stat} > F_{\alpha/2, m-1, n-1}$	

Test for Equivalence of Variance

Null hypothesis: $H_0 : \sigma_1^2 = \sigma_2^2$

same

Test statistic value:

$$F = s_1^2 / s_2^2$$

ratio of variances

Alt Hypothesis

Rejection Region

p-value:

$$H_a : \sigma_1^2 > \sigma_2^2$$

$$F_{stat} > F_{\alpha, m-1, n-1}$$

$$P(F_{m-1, n-1} > F_{stat})$$

$$H_a : \sigma_1^2 < \sigma_2^2$$

$$F_{stat} < F_{1-\alpha, m-1, n-1}$$

$$P(F_{m-1, n-1} < F_{stat})$$

$$H_a : \sigma_1^2 \neq \sigma_2^2$$

$$F_{stat} < F_{1-\alpha/2, m-1, n-1}$$

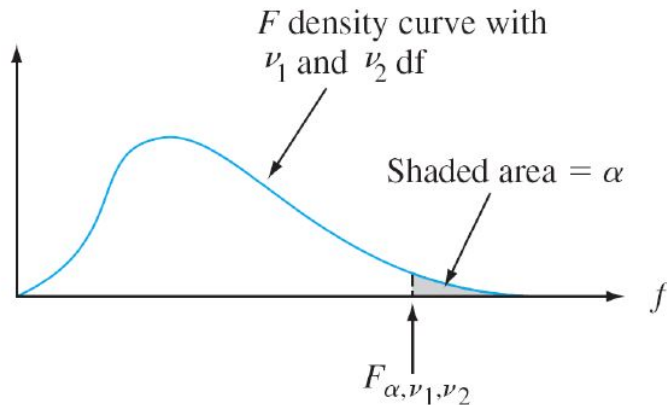
(OR)

$$\text{OR } F_{stat} > F_{\alpha/2, m-1, n-1}$$

Stats. F. ppf
.cdf

Test for Equivalence of Variance

Pictured: a typical F density function. When this thing took a value far from 1, we could conclude that the *ratio* being calculated had significantly different numerator from denominator. This is how we compared two variances.



The F-test

We use F statistics to compare variances. One way to compare linear models is to compare the variance in Y to the variance of your model: if your model is capturing a lot of the variance in Y , it's doing well!

In MLR we test the hypothesis:

var. (Y): $\frac{\sum (Y_i - \bar{Y})^2}{n-1}$

MLR w/ different choices of X/columns

SSE of \rightarrow model with 0 slope

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0 \rightarrow \text{any slopes}$$

which says that there is no useful linear relationship between y and any of the p predictors. We test against:

$$H_a : \text{any of the } B'_j\text{'s are nonzero.}$$

Compare that to SSE: $\sum (Y_i - \hat{Y}_i)^2$

We could test each separately, but we would be committing the multiple comparisons fallacy. A better test is a joint test, and is based on a statistic that has an F distribution when H_0 is true.

The Full F

Null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0 : \text{we don't need a}$$

Alternative hypothesis:

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

model: None of the
columns help us.

Test statistic value:

$$F = \frac{SSR/(p+1) \text{ (model)}}{SST/(n-p-1) \text{ (SST \& variance of } Y)}$$

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

The Partial F

Comparing variances also gives us another way - besides just adjusted R^2 - to compare between models.

Idea: compare the amount of variance captured by the larger model to the smaller model. If they're significantly different, we know the larger model is "adding" lots of information!

As a hypothesis, this means testing that the parameters that are different between models are zero.

The Partial F

Suppose we wanted to know whether *some* subset of model parameters were zero:

Null:

Alternative:

The Partial F

Suppose we wanted to know whether *some* subset of model parameters were zero:

Full model: $\underline{Y} = \beta_0 + \beta_1 \underline{X_1} + \dots + \beta_4 \underline{X_4}$ \hookleftarrow use x_1, x_2, x_3, x_4

Reduced Model: $\underline{Y} = \beta_0 + \beta_2 \underline{X_2} + \beta_4 \underline{X_4}$ \hookleftarrow only use x_2 & x_4 ; exclude x_1, x_3

Null:

Alternative:

The Partial F

Suppose we wanted to know whether *some* subset of model parameters were zero:

Full model: $\underline{Y} = \beta_0 + \beta_1 \underline{X_1} + \dots + \beta_4 \underline{X_4}$

Reduced Model: $\underline{Y} = \beta_0 + \beta_2 \underline{X_2} + \beta_4 \underline{X_4}$

Null: $H_0 : \beta_1 = \beta_3 = 0$ *⊂ assume smaller is better; assume we don't need β_1, β_3 .*

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

Suppose we wanted to know whether *some* subset of model parameters were zero:

Full model: $\underline{Y} = \beta_0 + \beta_1 \underline{X_1} + \dots \beta_4 \underline{X_4}$

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

every thing

The test is carried out by fitting both the full and reduced models.

↳ sets of columns/x-values
to exclude

Because the full model contains not only the predictors of the reduced model but also some extra predictors, it should fit the data at least as well as the reduced model.

That is, if we let SSE_{full} be the sum of squared residuals for the full model and SSE_{red} be the corresponding sum for the reduced model, then

$$\frac{\text{total}}{\text{prior}} < \frac{\text{reduced}}{\text{error}}$$

The Partial F

The test is carried out by fitting both the full and reduced models.

Because the full model contains not only the predictors of the reduced model but also some extra predictors, it should fit the data at least as well as the reduced model.

That is, if we let $\underline{SSE_{full}}$ be the sum of squared residuals for the full model and $\underline{SSE_{red}}$ be the corresponding sum for the reduced model, then $\underline{SSE_{full} < SSE_{red}}$

The Partial F

Intuitively, if _____ is a great deal smaller than _____, the full model provides a much better fit than the reduced model; the appropriate test statistic should then depend on the reduction _____ in unexplained variation.

Test statistic value:

Rejection region:

The Partial F

Intuitively, if $\underline{SSE_{full}}$ is a great deal smaller than $\underline{SSE_{red}}$, the full model provides a much better fit than the reduced model; the appropriate test statistic should then depend on the reduction $\underline{SSE_{red} - SSE_{full}}$ in unexplained variation.

Test statistic value:

$$F = \frac{(SSE_{red} - SSE_{full}) / (p - k)}{SSE_{full} / (n - (p + 1))}$$

improvement in
SSE

Rejection region:

Component to
S.S. full model

The Partial F

Intuitively, if $\underline{SSE_{full}}$ is a great deal smaller than $\underline{SSE_{red}}$, the full model provides a much better fit than the reduced model; the appropriate test statistic should then depend on the reduction $\underline{SSE_{red} - SSE_{full}}$ in unexplained variation.

Test statistic value:

$$F = \frac{(SSE_{red} - SSE_{full}) / (p - k)}{SSE_{full} / (n - (p + 1))}$$

Rejection region: $f \geq F_{\alpha, p-k, n-(p+1)}$

or use p-value: comes from summary table of
OLS (y, x) . f.t.t. summary

The Partial F

Summary:

The F distribution also lends itself to pretty nuanced comparisons. Instead of just comparing the variance of our model to the variance of Y , we can compare variances of different models.

Null:

Alternative:

The Partial F

Summary:

The F distribution also lends itself to pretty nuanced comparisons. Instead of just comparing the variance of our model to the variance of Y , we can compare variances of different models.

Full model: $\underline{Y} = \beta_0 + \beta_1 \underline{X_1} + \dots \beta_4 \underline{X_4}$

Reduced Model: $\underline{Y} = \beta_0 + \beta_2 \underline{X_2} + \beta_4 \underline{X_4}$

Null:

Alternative:

The Partial F

Summary:

The F distribution also lends itself to pretty nuanced comparisons. Instead of just comparing the variance of our model to the variance of Y , we can compare variances of different models.

Full model: $\underline{Y} = \beta_0 + \beta_1 \underline{X_1} + \dots \beta_4 \underline{X_4}$

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 .

The Partial F

Summary:

The F distribution also lends itself to pretty nuanced comparisons. Instead of just comparing the variance of our model to the variance of Y , we can compare variances of different models.

Full model: $\underline{Y} = \beta_0 + \beta_1 \underline{X_1} + \dots \beta_4 \underline{X_4}$

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

MLR so far:
1) Y continuous

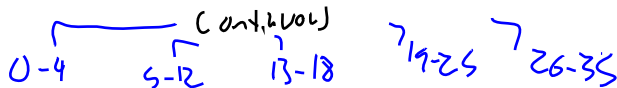
2) continuous X

The F distribution is commonly used for a type of analysis that overlaps with MLR: ANOVAs of **Analysis of Variance**.

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

1. Binary outcomes
2. Categorical outcomes
3. Artificial stratifications of the data

↳ age bracket



X 's are discrete:
binary (yes/no)
categories

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

We find the following:

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

9 people

ANOVA

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{3} + \frac{s_y^2}{3}}}$$

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 = "

F-test: test
"do groups matter"

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.

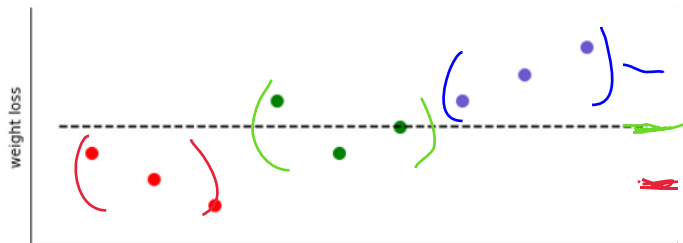
The one-way ANOVA model can be used to test the null hypothesis:

$H_0: \mu_A = \mu_B = \mu_C \dots$ for all groups/categories . *different LEL are 0.*

One-Way ANOVA

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.

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

Global mean \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

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:

data vs. \sum global mean

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

data vs. \sum local mean

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

6 vs. 4

One-Way ANOVA

1. **Total** sum of squares:

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

$\underbrace{\sum_{i=1}^I}_{\text{groups}} \underbrace{\sum_{j=1}^{n_i}}_{\text{points}} \uparrow \uparrow \text{global mean}$

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

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

$\uparrow \uparrow \text{local mean}$

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

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

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

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

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

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

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

$$\begin{aligned} \text{SSB} &= \sum_{i=1}^I \sum_{j=1}^{n_i} (\bar{y}_i - \bar{\bar{y}})^2 = \sum_{i=1}^I n_i (\bar{y}_i - \bar{\bar{y}})^2 \\ &= 3(2 - 4)^2 + 3(4 - 4)^2 + 3(6 - 4)^2 = 24 \end{aligned}$$

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

One-Way ANOVA

We have $\bar{y} = 4$ and $[\bar{y}_1, \bar{y}_2, \bar{y}_3] = [2, 4, 6]$

	Control	Diet A	Diet B	
0	4.5	3.2	5.4	5.6
1	4.5	2.2	3.1	6.6
2	4.5	1.2	4.4	7.6

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

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

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

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

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

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

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

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

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

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

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

Null hypothesis:

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

Alternative hypothesis:

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

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:

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

Test statistic value:

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

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

What are our assumptions?

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:

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

Test statistic value:

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

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

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

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

ANOVA tables

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

of SIPS
n=9

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

n = # of SIPS

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

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

ratio? $F \rightarrow$ tests

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		

Déjà vu

sm. OLS(Y, X)

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

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

Categories!

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: $y = \text{Control} + \text{effects of group} + \text{errors}$

intercept

→ slope: group matters!

So a data point in group 1 would look like:

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

+ β_A + ε

and a data point in group 2 would look like:

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

$\beta_0 + \beta_B$

Linear ANOVAs

Control: $(x, y) \rightarrow ((0, 0), y)$

Handwritten notes:
 x_{ij} with arrow to $(0, 0)$
 A with arrow to $(0, 0)$
 x_{ij} with arrow to y
 B with arrow to y

More formally, we use dummy or indicator variables. Denote

$$x_{ij} = \begin{cases} 1 & \text{if data point } j \text{ is in group } i \\ 0 & \text{else} \end{cases} \quad (A, \#)$$

Handwritten notes:
 $A: (x, y) \rightarrow ((1, 0), y\#)$
 $B: (x, y) \rightarrow ((0, 1), \#)$

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

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

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
2	0	0
1	0	0
5	1	0
3	1	0
4	1	0
5	0	1
6	0	1
7	0	1

Control

Group A

2 columns

{ (0,0) for A,B
is neither
baseline

{ Group B

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.

$\mathbb{1}_{\text{event}}$ or I_{event} : 1 if event is true
0 else

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

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?

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

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 fully 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!

MLR WRapup

Our full MLR workflow. To pre-process, we add in columns for an intercept and any relevant indicator variables for categorical or binary x -values.

Step	Idea	Plots	Fixes
0	Explore	pairs	just enjoy
1	Candidates	pairplot	remove redundant
2	Linearity	Residuals vs X/Y	transform?
3	Normality	Histograms and QQ	transform?
4	Homoskedasticity	Component-Residual	hard
5	Uncorrelation	Component-Residual	hard
6	Outliers	Influence, Cook's	remove?

We loop steps 2-6 until we're done, deciding whether or not our model is improving after each iteration by using things like SSE , adjusted R^2 , and F -tests to compare between models.

Final steps: Re-run the model

Don't make too many changes at a time. Fix one or two things and re-run the model each time. What fixes one thing can and will lead to different points becoming outliers and often changes to every one of the diagnostics above.

Don't rush a regression problem. It's most a gradual process of not throwing out too much and making sure assumptions are met in the final model.

Extra plots

Python has some automated functions for most of these plots! In one of the in class notebooks, we make a `FIT_AND_RES` function that did residuals plots against a single predictor. `statsmodels` includes a few ways to automate some other commons plots. Given a model from `MODEL=SM.OLS(Y,X).FIT()`

1. pairs plots, via `SEABORN.PAIRPLOT`. **Visually** assess related X values, then check them numerically with `SM.STATS.OUTLIERS_INFLUENCE.VARIANCE_INFLATION_FACTOR(X,I)`
2. QQ plots to determine normality, via `SM.QQPLOT(MODEL.RESID, STATS.T, DDOF)`. The the QQ plot isn't visually close to a straight line, it may mean that the errors are not normal. (it's plotting sample quantiles against theoretical t-critical value quantiles)
3. Component-residual plots. For a plot of X_i versus *resid*, you can use `STATSMODELS.PLOT_REGRESS_EXOG(MODEL, I)`. Here we look for whether transformations or polynomials of *that* X should be used.
4. `SM.GRAPHICS.INFLUENCE_PLOT(SLR, CRITERION="COOKS")`. *Discuss* and **maybe** remove offenders

Final steps: Reporting your results

Your final writeup should include the following:

1. All terms in your final model, their estimates, and their confidence intervals. You may (probably should) also include a sentence interpreting each one and whether or not you find that result intuitive/reasonable. Keep in mind that β_j is the effect of the j th predictor given the inclusion of all of your other predictors.
2. The F -test p-value and whether or not you reject the associated hypothesis. What does this mean?
3. Discuss whether your final model includes any insignificant predictors (e.g. the t -tests for those predictors would not have you reject H_0 .) Why are they in your model?
4. How useful is your model? Discuss R^2 , and any questions about prediction.

Final steps: Choosing between Models

Let's say you've "tuned" 2 or 3 different models with different subsets of predictors included. How do we choose between them?

1. Best adjusted R^2
2. Best AIC or BIC.
3. Lowest Standard Error (least SSE)?
4. Only significant terms included?

Level 2: Automating the Choice between Models

If you want to do a quick search on the optimization between models (without stopping between each to tune, there are 3 common schema)

1. Test all subsets (how many computations is this?)
2. Forward inclusion (add the best "missing" term each step until there's no longer anything useful to add)
3. Backward inclusion