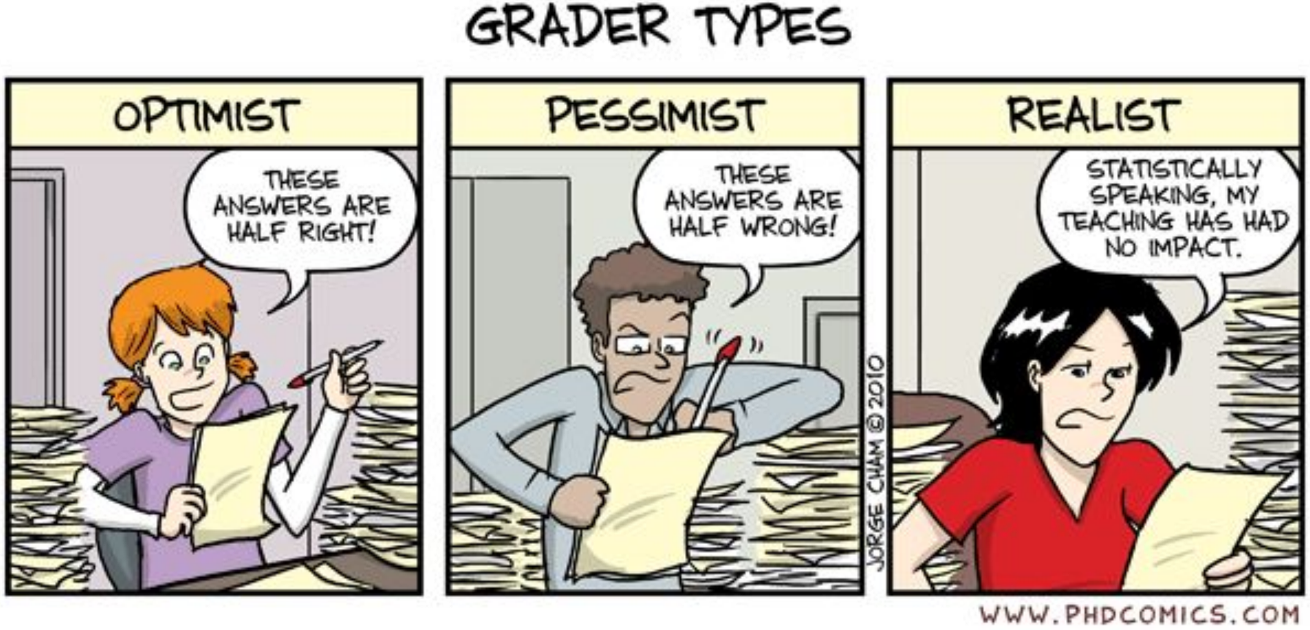
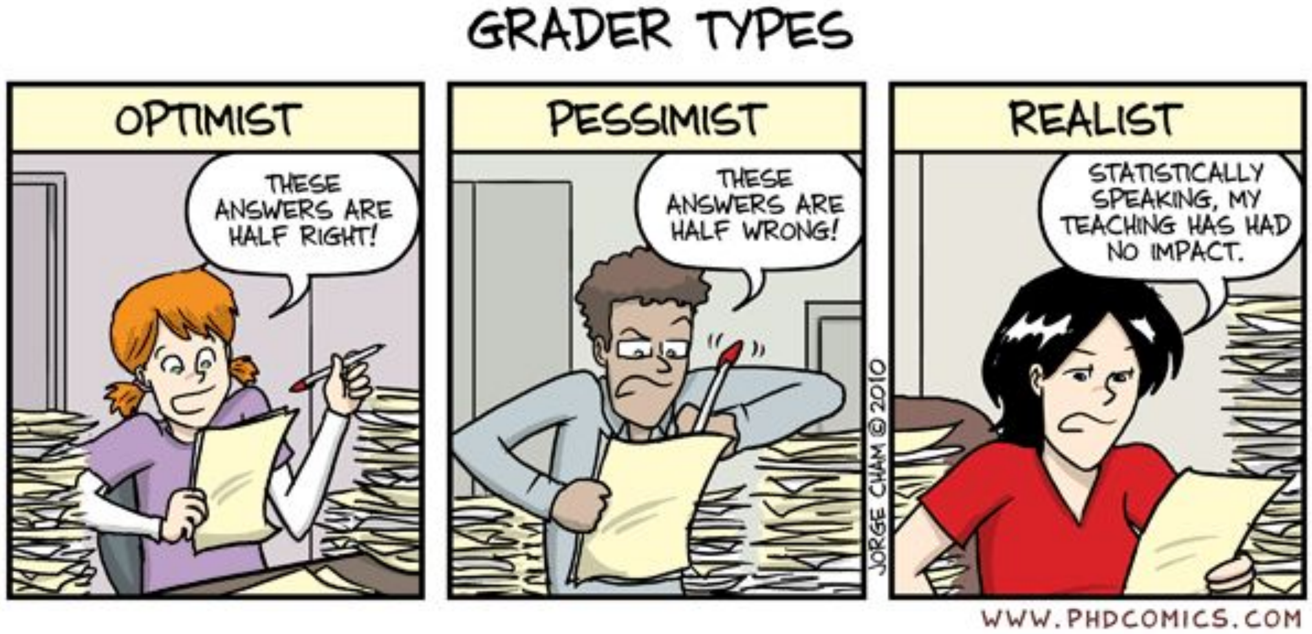
Lecture 24: Analysis of Variance (ANOVA)



Announcements and reminders

- HW 5 due Friday at 5 PM



Previously on CSCI 3022…

Given data (x_{i1}, x_{i2}, … , x_{ip}, y_i), for i = 1, 2, …, n, fit a MLR model of the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i, \quad \text{where each } \epsilon_i \sim N(0, \sigma^2)$$

We can test if any of the features are important using an **F-test**:

$$F = \frac{(SST - SSE)/p}{SSE/(n-p-1)} \qquad SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2 \qquad SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

The F-statistic follows an F-distribution:

Rejection region: $F \geq F_{\alpha,p,n-p-1}$

p-value: 1 - stats.f.cdf(F, p, n-p-1)

Comparing multiple means

We’re often interesting in comparing the means of a response from different groups

Example: S’pose we are doing a study on the effect of diet on weight-loss.
We have three different groups in the study:

$\mu_A \neq \mu_{control}$, $\mu_B \neq \mu_A$? $\mu_B \neq \mu_{control}$

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

Is there any evident of these mean are not equal to another

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

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

Comparing multiple means

We’re often interesting in comparing the means of a response from different groups

Example: S’pose we are doing a study on the effect of diet on weight-loss.
We have three different groups in the study:

- 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 results:

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

Comparing multiple means

We're often interesting in comparing the means of a response from different groups

Example: S'pose we are doing a study on the effect of diet on weight-loss.
We have three different groups in the study:

- 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 results:

Question: Are the means of the different groups all the same?

And: What would we do if there were only two groups?

t - test

Comparing multiple means

We're often interesting in comparing the means of a response from different groups

Example: S'pose we are doing a study on the effect of diet on weight-loss.
We have three different groups in the study:

- 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 results:

Question: Are the means of the different groups all the same?

But also: Why would a t- or z- test be problematic if we have **many** different groups?

Analysis of variance

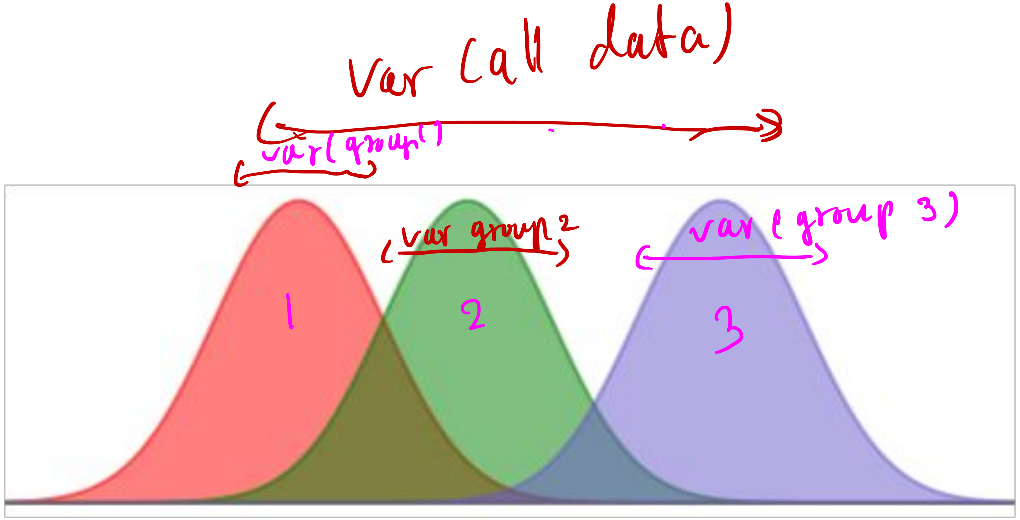
We can answer the question “Are any of the means different?” by using a procedure called **analysis of variance**, or **ANOVA**

The idea: Look at where the **variance in the data** comes from

Analysis of variance

We can answer the question “Are any of the means different?” by using a procedure called **analysis of variance**, or **ANOVA**

The idea: Look at where the **variance in the data** comes from



Analysis of variance

We can answer the question “Are any of the means different?” by using a procedure called **analysis of variance**, or **ANOVA**

index groups as $i = 1, 2, \ldots, I$

The idea: Look at where the **variance in the data** comes from

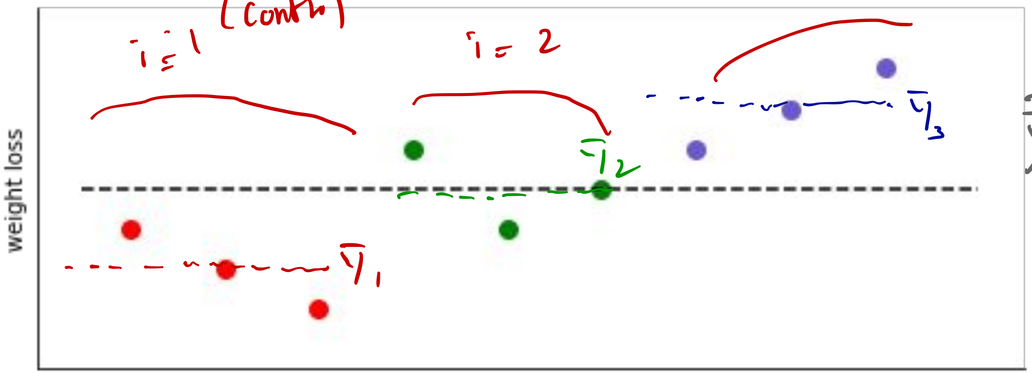
Grand mean = $\bar{\bar{Y}}$

Group means = $\bar{Y}_i$

\bar{Y}_i
\overline{Y}

(control) (diet 2) $i = 3$ (diet B)

$I = 3$
group: w/ has A, B



$$\bar{\bar{Y}} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} Y_{ij}}{\#\ \text{data pts}}$$

Analysis of variance

S'pose we have I groups that we want to compare, each with n_i data points ($i = 1, 2, \ldots, I$)

- The **grand mean** is the sample mean of all responses:

$$\bar{Y} = \frac{1}{\text{data pts}} \sum (\text{all data pts}) = \frac{1}{9}(6 + 4 + 2)$$

- The **group means** are the sample means within each group:

$$\bar{Y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \longrightarrow \bar{Y}_1 = \frac{1}{3} \sum_{j=1}^{3} Y_{1j} = \frac{1}{3}[3 + 2 + 1]$$

$$= 2$$

$$\bar{Y}_2 = \frac{1}{3}[5 + 3 + 4] = 4$$

$$\bar{Y}_3 = \frac{1}{3}[5 + 6 + 7] = 6$$

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

$n_1 = 3$

It's the *variances*, stupid!

Where does the total variation in the data come from?

I = # of group

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

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

A minor mathematical miracle: $SST = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}})^2 = \sum \sum ((y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{\bar{y}}))^2$

$$SST = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \left[(y_{ij} - \bar{y}_i)^2 + (\bar{y}_i - \bar{\bar{y}})^2 \right] = \underbrace{\sum_{i=0}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}_{SSW} + \underbrace{\sum_{i=1}^{I} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{\bar{y}})^2}_{SSB}$$

$\rightarrow SST = SSW + SSB$

It’s the *variances*, stupid!

Where does the total variation in the data come from?

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

A helpful decomposition:

$$y_{ij} - \overline{\overline{y}} = \underbrace{(y_{ij} - \overline{y_i})}_{\text{within group}} + \underbrace{(\overline{y_i} - \overline{\overline{y}})}_{\text{between group}}$$

A minor mathematical miracle:

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

$$= SSW + SSB$$

The one-way ANOVA model

	Control	Diet A	Diet B
0	3	5	5
1	2	3	6
2	1	4	7
$\bar{Y}_i$	2	4	6
$\bar{\bar{Y}}$	4		

Let's compute the variances (.... or sum of squares) for our data!

- The **between groups** sum of squares is: $SSB = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{\bar{y}})^2$

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

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

$= 3(4) + 3(4) = \boxed{24}$

- The **within groups** sum of squares is:

$(n-1)\,\mathrm{var}(y_{ij})$

$SSW = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 = [(3-2)^2 + (2-2)^2 + (1-2)^2] + [(5-4)^2 + (3-4)^2 + (4-4)^2]$

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

$= \boxed{6}$

- The **total** sum of squares is:

$SST = \sum\sum(y_{ij} - \bar{\bar{y}})^2 = SSW + SSB = 24 + 6 = 30$

The one-way ANOVA model

Let's compute the variances (.... or sum of squares) for our data!

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

- The **between groups** sum of squares is:

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

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

- The **within groups** sum of squares is:

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

$$= [(3-2)^2 + (2-2)^2 + (1-2)^2] + [(5-4)^2 + (3-4)^2 + (4-4)^2] +$$
$$[(5-6)^2 + (6-6)^2 + (7-6)^2] \; = \; 6$$

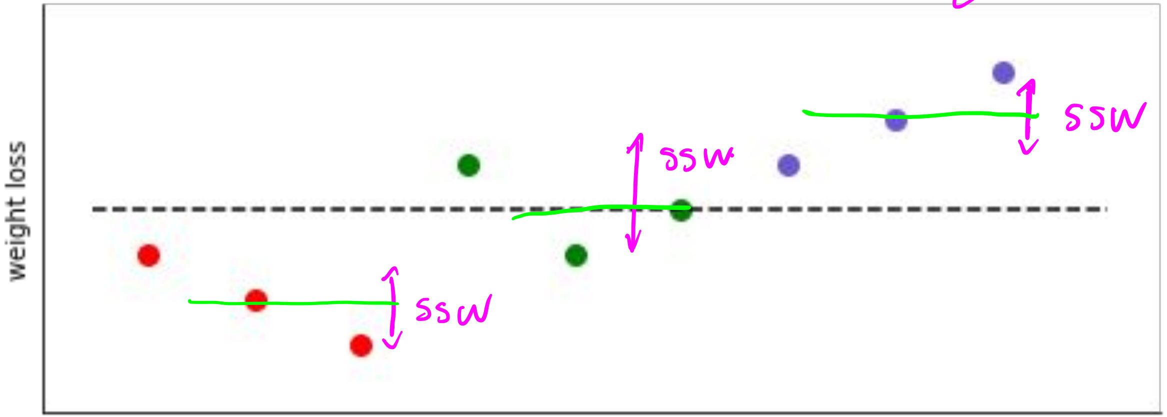
- The **total** sum of squares is:

$$SST = SSB + SSW = 24 + 6 = 30$$

The one-way ANOVA model

Compare these results to the original picture:

want to compare SSW & SSB
if SSB >> SSW than
it feels like there's probability
a diff b/w group



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

$$MLR \Rightarrow F = \frac{(SSE_{red} - SSE_{full})/df_{num}}{SSE_{full}/df_{num}}$$

The one-way ANOVA model

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} = ________

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

"data" ~ $\bar{y}_i$ ← 3 of these (general: I)
estimate $\bar{\bar{y}}$

- The **within groups** degrees of freedom is: SSW_{df} = ________

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

$N = n_1 + n_2 + \dots + n_I$ = total # data pts

- So for our example, we have: SSB_{df} = __2__ and SSW_{df} = __6__

$N = 9$ data point total

$I = 3$ groups

The one-way ANOVA model

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: $SSB_{df} = I - 1$

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

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

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

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

We <3 hypothesis testing

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

H_0: ______________________

H_1: ______________________

Our test statistic will be: F = ______________________ $\sim F_{I-1,N-I}$

Rejection region:

p-value:

We <3 hypothesis testing

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

H$_0$: $\mu_1 = \mu_2 = \ldots = \mu_I$

H$_1$: $\mu_i \neq \mu_j \quad \text{for some pair} \quad i, j$

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

stats.f.ppf(1-α, dfn= I -1, dfd = N-I)

Rejection region: $F \geq F_{\alpha,I-1,N-I}$

p-value: 1 - stats.f.cdf(*F*, *I*-1, *N*-*I*)

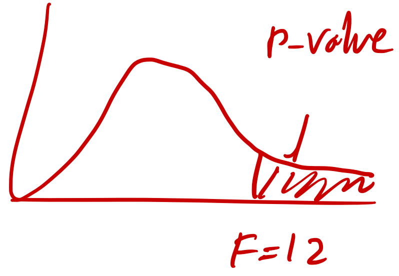
The ANOVA table

It is common practice to organize all computations into an **ANOVA table**

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

$$F = \frac{SSB/df_{SSB}}{SSW/df_{SSW}}$$

ANOVA	SS	DF	$\boxed{\text{SS/DF}}$	F
between	24	2	12	12
within	6	6	1	0.008
total	30			



p-value = 1 - stats.f.cdf(12, dfn = 2, dfd = 6)

= 0.008

The ANOVA table

It is common practice to organize all computations into an **ANOVA table**

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

p_value = 1 - stats.f.cdf(F, dfnum, dfden)

ANOVA	SS	DF	SS/DF	F
between	24	2	12	12
within	6	6	1	p = 0.008
total	30	8		

doesn't tell us

$p > \alpha$? fail to reject

$p < \alpha$? → reject H_0

if $\alpha = 0.01$, we reject H_0 & could be there is evidence of some difference in group means

ANOVA as a multiple linear regression

Interestingly, there is a close relationship between **One-Way ANOVA** and **MLR**

S'pose you have I groups that you want to compare.
A random sample of size n_i is taken from the i^{th} group. Then…

- Choose one group as the control

I groups

-1

- 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, and

$$x_{ij} = \begin{cases} 1 & \text{if } j^{th} \text{ response is from } i^{th} \text{ group} \\ 0 & \text{otherwise} \end{cases}$$

ANOVA as a multiple linear regression $Y_{ij} = \mu_0 + \tau_1 x_{1j} + \tau_2 x_{2j} + \varepsilon_{ij}$

Interestingly, there is a close relationship between **One-Way ANOVA** and **MLR**

S'pose you have I groups that you want to compare.
A random sample of size n_i is taken from the i^{th} group. Then…

$$y_{ij} = \mu_0 + \tau_1 x_{1j} + \tau_2 x_{2j} + \ldots + \tau_{I-1} x_{I-1,j} + \epsilon_{ij}$$

$$x_{ij} = \begin{cases} 1 & \text{if } j^{th} \text{ response is from } i^{th} \text{ group} \\ 0 & \text{otherwise} \end{cases}$$

fit on MLR model with feature

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

control

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

Interestingly, there is a close relationship between **One-Way ANOVA** and **MLR**

S'pose you have I groups that you want to compare.
A random sample of size n_i is taken from the i^{th} group. Then…

$$y_{ij} = \underline{\mu_0} + \underline{\tau_1} x_{1j} + \underline{\tau_2} x_{2j} + \ldots + \tau_{I-1} x_{I-1,j} + \epsilon_{ij}$$

MLR gives us μ_0, τ_1, τ_2

$$x_{ij} = \begin{cases} 1 & \text{if } j^{th} \text{ response is from } i^{th} \text{ group} \\ 0 & \text{otherwise} \end{cases}$$

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

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

ANOVA as a multiple linear regression

Interestingly, there is a close relationship between **One-Way ANOVA** and **MLR**

S'pose you have I groups that you want to compare.
A random sample of size n_i is taken from the i^{th} group. Then…

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

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

Mean response for control: control : $x_{1j} = x_{2j} = 0$

$\rightarrow Y_{ij} = \mu_0$

Mean response for Diet A: group 1 : $x_{1j} = 1$ & $x_{2j} = 0$

$\rightarrow Y_{1j} = \mu_0 + \tau_1$

↳ effect for group 1

Mean response for Diet B:

group 2 : $x_{1j} = 0$ & $x_{2j} = 1$

$\rightarrow Y_{2j} = \mu_0 + \tau_2$

ANOVA as a multiple linear regression

Interestingly, there is a close relationship between **One-Way ANOVA** and **MLR**

S'pose you have I groups that you want to compare.
A random sample of size n_i is taken from the i^{th} group. Then…

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

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

Interestingly, there is a close relationship between **One-Way ANOVA** and **MLR**

S'pose you have I groups that you want to compare.
A random sample of size n_i is taken from the i^{th} group. Then… $y_{ij} = \mu_0 + \tau_1 x_{1j} + \tau_2 x_{2j}$

Control: μ_0
Diet A: $\mu_1 = \mu_0 + \tau_1$
Diet B: $\mu_2 = \mu_0 + \tau_2$ $\leftarrow$ τ_1 and τ_2 are the **treatment effects** for the two diets

ANOVA as a multiple linear regression

Interestingly, there is a close relationship between **One-Way ANOVA** and **MLR**

S'pose you have I groups that you want to compare.
A random sample of size n_i is taken from the i^{th} group. Then…

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

Control: μ_0
Diet A: $\mu_1 = \mu_0 + \tau_1$
Diet B: $\mu_2 = \mu_0 + \tau_2$ ← τ_1 and τ_2 are the **treatment effects** for the two diets

$H_0(MLR) \rightarrow H_0(ANOVA)$
$H_0(ANOVA) \Leftrightarrow H_0(MLR)$

MLR F-test: $H_0: \tau_1 = \tau_2 = \dots = \tau_p = 0$
H_1: at least one of $\tau_k \neq 0$

ANOVA equivalent: $H_0: \mu_0 = \mu_1 = \mu_2$
H_1: at least one pair $\mu_i \neq \mu_j$ $(i \neq j)$

Tukey's honest significance test

S'pose we determine that some of the means are different

How can we tell which ones?

→ Tukey's HST (**aka** Tukey's Range Test **aka** Tukey's Honest Significant Difference (HSD))

- Hypothesis test for pairwise comparison of means
  - It's just lots of pairwise tests using what's called the **studentized range distribution** ← very cool, but only FYI
- Adjusts so that prob of making a Type I error over **all** possible pairwise comparisons = α

  → Fixes Problem of Multiple Comparisons!

Tukey's honest significance test

S'pose we determine that some of the means are different

How can we tell which ones?

→ Tukey's HST (**aka** Tukey's Range Test **aka** Tukey's Honest Significant Difference (HSD))

- Hypothesis test for pairwise comparison of means (it's just lots of pairwise tests)
  - It's just lots of pairwise tests using what's called the **studentized range distribution** ← very cool, but only FYI
- Adjusts so that prob of making a Type I error over **all** possible pairwise comparisons = α

  → Fixes Problem of Multiple Comparisons!

100 tests each with

$P(\text{type I error}) = \alpha$

$\Rightarrow$ overall $P(\text{Type I error})$

$= 1-(1-\alpha)^{100}$

$1-(1-.05)^{100} = 1-.95^{100} = 0.94$

… What just happened?

… **ANOVA** = **AN**alysis **O**f **VA**riance just happened!

- How we can formally test for a difference among lots of group means
- Avoids needing to do lots of individual 2-sample tests
- Decomposes the total sum of squares into that attributable to **between-group** variation and **within-group** variation