

CSCI 3022

intro to data science with probability & statistics

November 28, 2018

Forward & Backward Selection

+

Analysis of Variance (ANOVA)

Arkaive! :-)



Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Last time on CSCI 3022:

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

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

- We can test if any of the features are important:

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

$$SST = \sum_{I=1}^n (y_i - \bar{y})^2$$

$$SSE = \sum_{I=1}^n (y_i - \hat{y}_i)^2$$

- The F-statistic follows an F-distribution
- Rejection Region: $F \geq F_{\alpha, p, n-p-1}$ p-value: $1 - \text{stats.f.cdf}(F, p, n-p-1)$

Is a Subset of Features Important?

- **Full Model:** $y = \beta_0 + \underline{\beta_1}x_1 + \underline{\beta_2}x_2 + \underline{\beta_3}x_3 + \underline{\beta_4}x_4$ (p=4 features in full model)
- **Reduced Model:** $y = \beta_0 + \underline{\beta_2}x_2 + \underline{\beta_4}x_4$ (k=2 features in reduced model)
- **Question:** Are the missing features important, or are we OK going with the reduced model?
- **Partial F-Test:** $H_0 : \beta_1 = \beta_3 = 0$
- Since the features in the reduced model are also in the full model, we expect the full model to perform at least as well as the reduced model.
- **Strategy:** Fit the Full and Reduced models. Determine if the difference in performance is real or due to just chance.

Is a Subset of Features Important?

- SSE_{full} = variation unexplained by the full model p features
- SSE_{red} = variation unexplained by the reduced model k features
($k < p$)

Intuitively, if SSE_{full} is much smaller than SSE_{red} , the full model fits the data much better than the reduced model. The appropriate test statistic should depend on the difference $SSE_{\text{red}} - SSE_{\text{full}}$ in unexplained variation.

- Test Statistic:

$$F = \frac{(SSE_{\text{red}} - SSE_{\text{full}})/(p - k)}{SSE_{\text{full}}/(n - p - 1)} \sim F_{p-k, n-p-1}$$

- Rejection Region:

$$F \geq F_{\alpha, p-k, n-p-1}$$

F... why even?

- Why compute the p-value for F-statistic when instead, we already have p-values for each of the covariates?
- Doing so would not be testing one hypothesis, but rather p hypotheses!
- At $\alpha=0.05$, how many p values do we expect to be significant if the null hypothesis is, in fact, true?

Suppose 100 features. $(0.05 \cdot 100) = 5$ false pos.
Features simply by chance!

In [27]:	1	model.summary()				
Out[27]: OLS Regression Results						
Dep. Variable:	sales	R-squared: 0.897				
Model:	OLS	Adj. R-squared: 0.896				
Method:	Least Squares	F-statistic: 570.3				
Date:	Tue, 28 Nov 2017	Prob (F-statistic): 1.58e-96				
Time:	20:28:02	Log-Likelihood: -386.18				
No. Observations:	200	AIC: 780.4				
Df Residuals:	196	BIC: 793.6				
Df Model:	3	?				
Covariance Type:	nonrobust	?				
	coef	std err	t	P> t	[0.025	0.975]
const	2.9389	0.312	9.422	0.000	2.324	3.554
tv	0.0458	0.001	32.809	0.000	0.043	0.049
radio	0.1885	0.009	21.893	0.000	0.172	0.206
news	-0.0010	0.006	-0.177	0.860	-0.013	0.011

The road to R² for MLR

- Just as with simple regression, the error sum of squares is:

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\hat{r}^2 = \frac{SSE}{n - (p+1)} = \frac{SSE}{n - p - 1}$$

Note: you may see SSE written as RSS: "residual sum of squares"

- It is again interpreted as a measure of how much variation in the observed y values is not explained by (not attributed to) the model relationship.
- The number of df associated with SSE is $n - (p+1)$ because $p+1$ df are lost in estimating the $p+1$ β coefficients.

The road to R^2

- Just as before, the **total sum of squares** is:

$$SST = \sum_{i=1}^n (\gamma_i - \bar{\gamma})^2 \quad df: n - 1$$

- And the **sum of squared errors** is:

$$SSE = \sum_{i=1}^n (\gamma_i - \hat{\gamma}_i)^2 \quad df: n - p - 1$$

- Then the coefficient of multiple determination R^2 is:

$$R^2 = 1 - \frac{SSE}{SST} \quad (SLR)$$

- It is interpreted in the same way as before. (Do you remember?)

Hacking R^2

Unfortunately, there is a problem with R^2 : Its value can be inflated by adding lots of predictors into the model even if most of these predictors are frivolous!

Hacking R^2

- For example, suppose y is the sale price of a house. Then:
- Sensible predictors include
 x_1 = the interior size of the house,
 x_2 = the size of the lot on which the house sits,
 x_3 = the number of bedrooms,
 x_4 = the number of bathrooms, and
 x_5 = the house's age.
- But now suppose we add in
 x_6 = the diameter of the doorknob on the coat closet,
 x_7 = the thickness of the cutting board in the kitchen,
 x_8 = the thickness of the patio slab.

Adjusted R²

- The objective in multiple regression is not simply to explain most of the observed y variation, but to do so using a model with relatively few predictors that are easily interpreted.
- It is thus desirable to adjust R² to take account of the size of the model:

$$R^2 = 1 - \frac{SSE}{SST}$$

$$R_a^2 = 1 - \frac{\frac{SSE}{df_{SSE}}}{\frac{SST}{df_{SST}}} = 1 - \frac{\frac{SSE}{(n-p-1)}}{\frac{SST}{(n-1)}}$$

↑
adjusted

$$= 1 - \frac{SSE(n-1)}{SST(n-p-1)}$$

Adjusted R²

- The objective in multiple regression is not simply to explain most of the observed y variation, but to do so using a model with relatively few predictors that are easily interpreted.
- It is thus desirable to adjust R² to take account of the size of the model:

$$R_a^2 = 1 - \frac{SSE/df_{SSE}}{SST/df_{SST}} = 1 - \frac{SSE/(n - p - 1)}{SST/(n - 1)}$$

Adjusted R²

In [27]:

```
1 model.summary()
```

Out[27]:

OLS Regression Results

Dep. Variable:	sales	R-squared:	0.897			
Model:	OLS	Adj. R-squared:	0.896			
Method:	Least Squares	F-statistic:	570.3			
Date:	Tue, 28 Nov 2017	Prob (F-statistic):	1.58e-96			
Time:	20:28:02	Log-Likelihood:	-386.18			
No. Observations:	200	AIC:	780.4			
Df Residuals:	196	BIC:	793.6			
Df Model:	3					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	2.9389	0.312	9.422	0.000	2.324	3.554
tv	0.0458	0.001	32.809	0.000	0.043	0.049
radio	0.1885	0.009	21.893	0.000	0.172	0.206
news	-0.0010	0.006	-0.177	0.860	-0.013	0.011

Deciding on important variables

- Suppose that we have 100 data points ($n=100$), but we have 200 different features ($p=200$). How can we learn which features are important and which are not?
- **Some options:**
 - Try all the possible combinations of features in models to see which gives the best fit.

Bad idea! Reason. 2^p diff. models.

$p = 3 \rightarrow 8$ models

$p = 30 \rightarrow 1,073,741,824$ models. yikes!

Deciding on important variables

- Suppose that we have 100 data points ($n=100$), but we have 200 different features ($p=200$). How can we learn which features are important and which are not?
- **Some options:**

- **Forward selection:**

baseline. 1. fit null model with an intercept but no predictors.

- baseline + 1 feature.* 2. fit p-SLRs, 1 for each feature. Choose the one that gives the lowest SSE. *keep that one! e.g. ft^2*
3. fit $p-1$ MLRs. Choose that which gives lowest SSE...
 $ft^2 + bathrooms$, $ft^2 + patio$, $ft^2 + dogs$, ...
4. repeat.

Deciding on important variables

- Suppose that we have 100 data points ($n=100$), but we have 200 different features ($p=200$). How can we learn which features are important and which are not?
- **Some options:**
 - **Backward selection:**
 1. Fit model with *all* predictors
 2. Remove the one with the largest p -value.
 3. Fit model with $p-1$ predictors.
 4. Remove the one with the largest p -value...

Quiz

1. **Advertising example.** I want to know if the set of {news,radio} have a slope that is significantly different from 0.
F-test. For a subset of features. "Partial" F-test. $H_0: \beta_{\text{news}} = \beta_{\text{radio}} = 0$
2. **Home prices example.** I have 1000 data points and 30 features. I want to learn the 10 most predictive and significant features.
Forward or Backward selection.
3. **Home prices example.** I have 100 data points and 200 features. I want to learn the 20 most predictive features.
Forward only. If $n < p$, use forward selection.
4. **Shark attacks example.** I have 50 shark attacks, and I have 20 features but they are *unlabeled*. I want to compute how well my model fits the data.

Use R_a^2 ← penalty for # features

Comparing multiple means

- We're often interested in comparing the means of a response from different groups
- **Example:** Suppose 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:

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

Comparing multiple means

- We're often interested in comparing the means of a response from different groups
- **Example:** Suppose 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?

What would we do if there were only two groups?

CI for $\mu_1 - \mu_2$? includes 0?

CI for μ_1 , CI for μ_2 ? non-overlapping?

t-test $H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

Comparing multiple means

- We're often interested in comparing the means of a response from different groups
- **Example:** Suppose 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?

Why would a t- or z-test be problematic if we had many different groups?

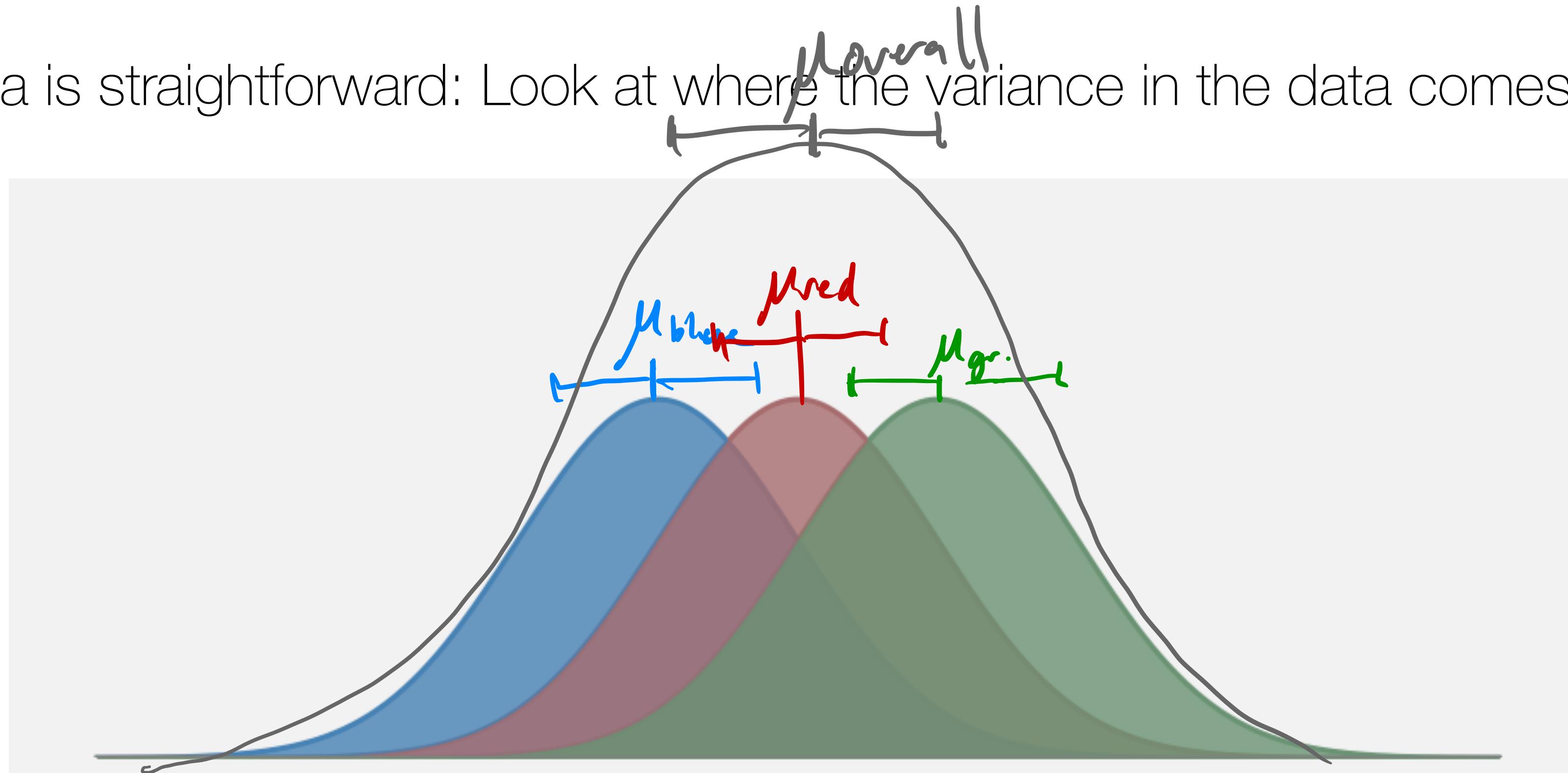
① expensive

k groups. $\Rightarrow \binom{k}{2}$ comparisons

② Multiple comparisons Problems!

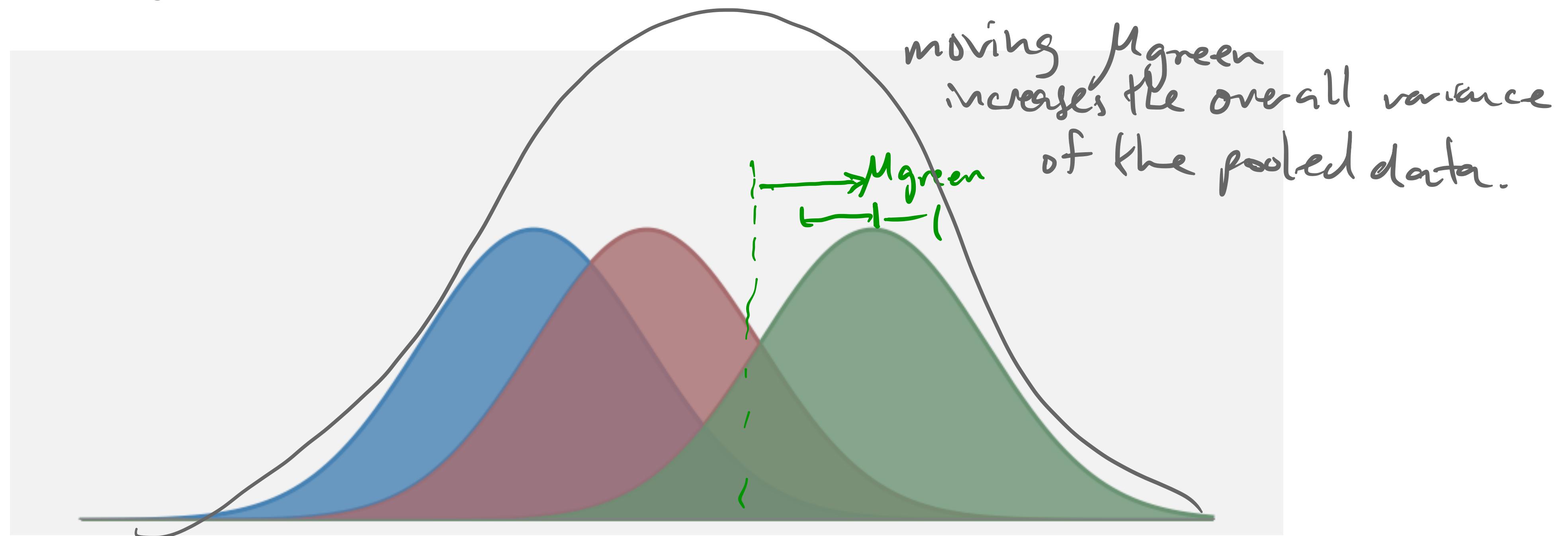
Analysis of variance

- We can answer the question “Are any of the means different?” using a procedure called analysis of variance, or **ANOVA** for short.
- The idea is straightforward: Look at where the variance in the data comes from.



Analysis of variance

- We can answer the question “Are any of the means different?” using a procedure called analysis of variance, or **ANOVA** for short.
- The idea is straightforward: Look at where the variance in the data comes from.



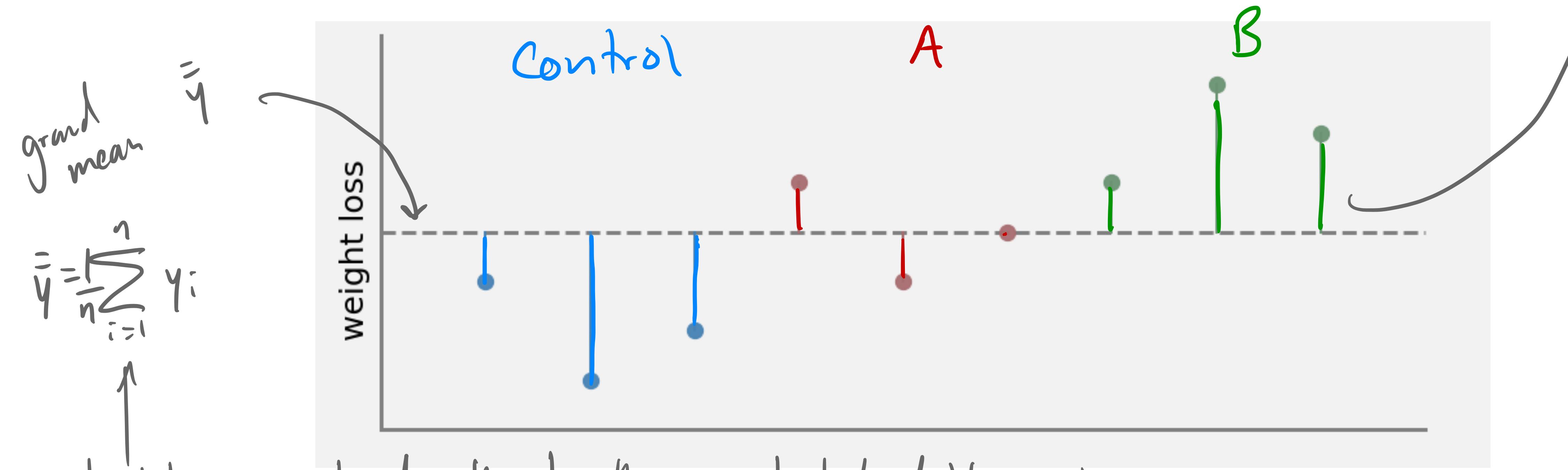
Analysis of variance

Overall variance:

$$\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

↑
"grand mean"

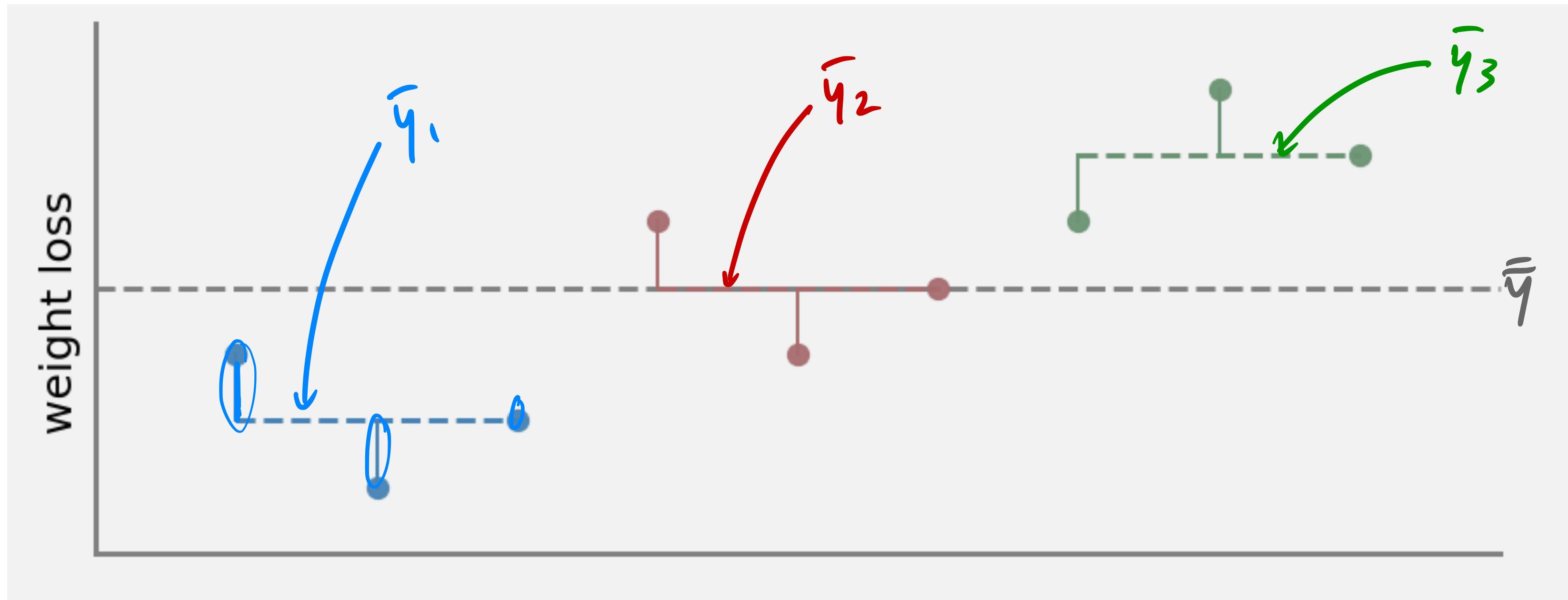
- We can answer the question "Are any of the means different?" using a procedure called analysis of variance, or **ANOVA** for short.
- The idea is straightforward: Look at where the variance in the data comes from.



calculation pretends that the group labels don't exist.

Analysis of variance

- We can answer the question “Are any of the means different?” using a procedure called analysis of variance, or **ANOVA** for short.
- The idea is straightforward: Look at where the variance in the data comes from.



The one-way ANOVA model

- Suppose that we have I groups that we want to compare, each with n_i data
- We model the relationship between responses and group means as follows:

$$y_{ij} = \mu_i + \epsilon_{ij}$$

group i

j th data point in the group

mean of group i

noise, error

$\epsilon_{ij} \sim N(0, \sigma^2)$

noise that is IID $N(0, \sigma^2)$

* Does not depend on which group we're looking at *

Assumptions:

- the responses are i.i.d. samples from normally distributed groups
- the variance of each group is the same

ANOVA is "pretty robust" to violations of these assumptions.

- non-normal errors.
- slightly non-equal variances.

The one-way ANOVA model

Let's compute some means!

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

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

$$= \frac{1}{9} [6 + 12 + 18] = \frac{1}{9} 36 = 4$$

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

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

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

$$\bar{y}_1 = \frac{1}{3} [3+2+1] = \frac{6}{3} = \boxed{2}$$

$$\bar{y}_3 = \frac{1}{3} [5+6+7] = \frac{18}{3} = \boxed{6}$$

$$\bar{y}_2 = \frac{1}{3} [5+3+4] = \frac{12}{3} = \boxed{4}$$

$$\begin{aligned} N &= n_1 + n_2 + n_3 \\ &= 3 + 3 + 3 = 9 \end{aligned}$$

note: $\bar{y} = \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2 + n_3 \bar{y}_3}{N}$

It's the variances, stupid

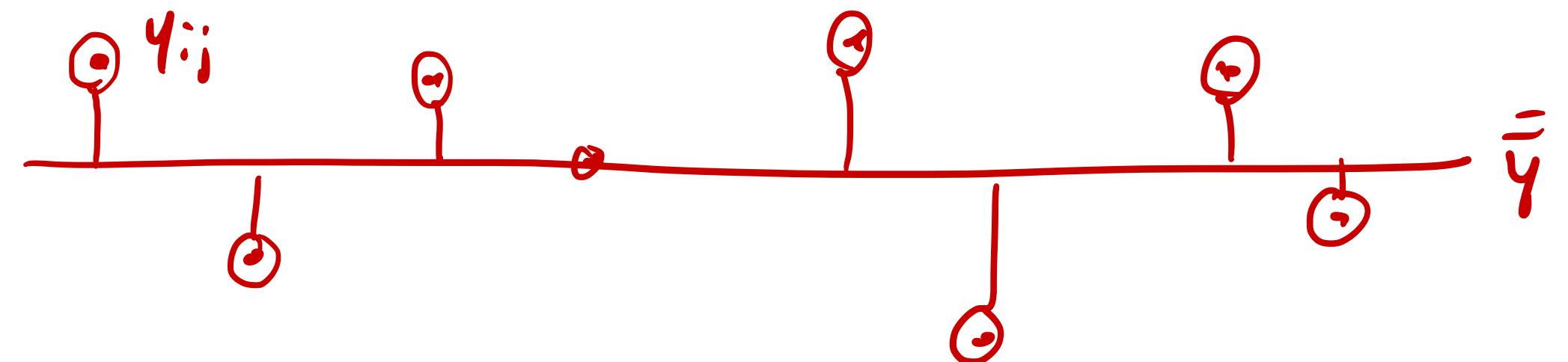
1992 Camille

cf "It's the economy, stupid."

- Where does the total variation in the data come from? Remember linear regression:

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

overall groups
within each group



- A helpful decomposition:

$$y_{ij} - \bar{y} + \bar{y}_i - \bar{y}_i = (\text{within-group deviation}) + (\text{between group deviation})$$

overall deviation

- Then, a minor (mathematical) miracle occurs:

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

SS Within
SS Between

The one-way ANOVA model

Let's compute some variances (or at least, sums of squares)!

- The **BETWEEN** group sum of squares is:

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

- The **WITHIN** group sum of squares is:

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

$$\begin{aligned} (3-2)^2 &= 1 & (5-4)^2 &= 1 & (5-6)^2 &= 1 \\ (2-2)^2 &= 0 & (3-4)^2 &= 1 & (6-6)^2 &= 0 \\ (1-2)^2 &= 1 & (4-4)^2 &= 0 & (7-6)^2 &= 1 \end{aligned} \quad \left. \right\} 6$$

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

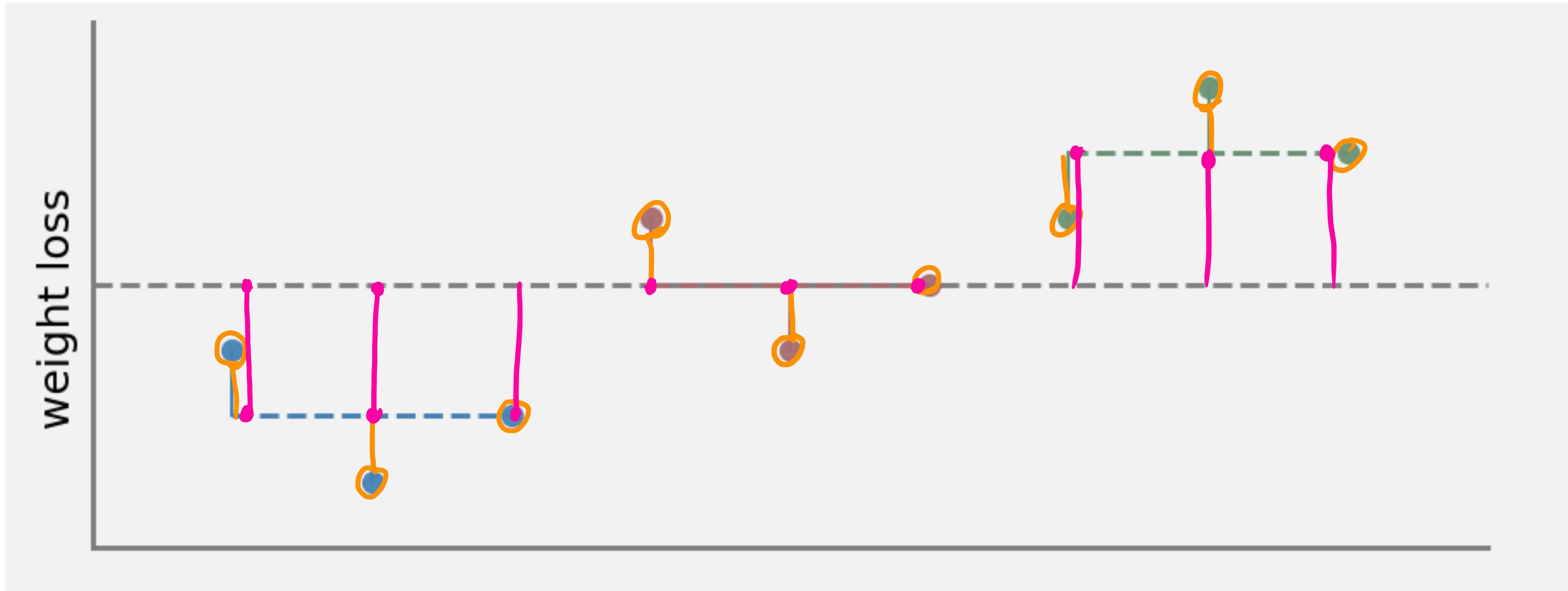
$$SST = SSB + SSW = 24 + 6 = 30$$

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

The one-way ANOVA model

$$SSB = 24$$
$$SSW = 6$$

- Compare these results to the original picture:



The one-way ANOVA model

What about degrees of freedom?

- The BETWEEN group degrees of freedom is (are?):

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

treat the \bar{y}_i like data

treat this as being computed from the data.

$$SSB_{df} = I - 1 = 3 - 1 = 2$$

- The WITHIN group degrees of freedom is (are?):

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

treat these as data

treat as computed from data

$$SSW_{df} = N - I = 9 - 3 = 6$$

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

A hypothesis test

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

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

$$H_1 : \mu_i \neq \mu_j \text{ for some } (i, j) \text{ pair, i.e., two of the group means are different.}$$

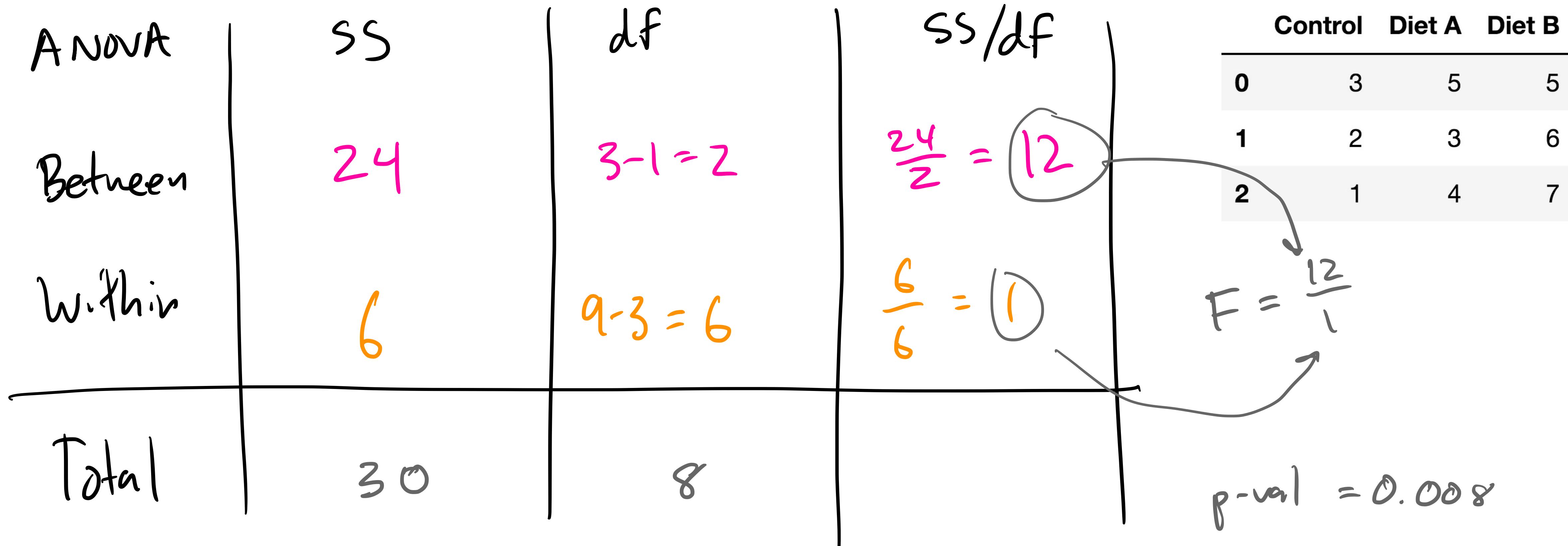
- Our test statistic will be:

$$F = \frac{SSB / df_{SSB}}{SSW / df_{SSW}} = \frac{SSB / (I-1)}{SSW / (N-I)} = \frac{SSB (N-I)}{SSW (I-1)}$$

Rejection Region Test $F \geq F_{\alpha, I-1, N-I}$ -or- p-val: 1-stats.f. cdf ($F, I-1, N-I$) compare to α .

The ANOVA Table

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



ANOVA as multiple linear regression

- Interestingly, there is a very close relationship between One-Way ANOVA and MLR!
- Suppose you have l groups that you want to compare. A random sample of size n_i is taken from the i^{th} group. Then

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

being part of A is feature X_1

B is feature X_2

y_{ij}	X_1	X_2
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 multiple linear regression

- Interestingly, there is a very close relationship between One-Way ANOVA and MLR!
- Suppose you have l groups that you want to compare. A random sample of size n_i is taken from the i^{th} group. Then

$$Y_{ij} = \mu_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_{ij} \quad \text{from } N(0, \sigma^2)$$

mean response for control group

$$X_1 = 0, \quad X_2 = 0 \quad \text{mean control} = \mu_0$$

mean response for diet A...

$$X_1 = 1, \quad X_2 = 0 \quad \text{mean A} = \mu_0 + \beta_1$$

$$\text{control: } \mu_0$$

$$\text{A: } \mu_0 + \beta_1$$

$$\text{B: } \mu_0 + \beta_2$$

etc. for B

$$\text{mean B: } \mu_0 + \beta_2$$