

**CSCI 3022**

# **intro to data science with probability & statistics**

November 28, 2018

Forward & Backward Selection  
+  
Analysis of Variance (ANOVA)

# 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:

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

- 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:**  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$  (p=4 features in full model)
- **Reduced Model:**  $y = \beta_0 + \beta_2 x_2 + \beta_4 x_4$  (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?

- $SSE_{\text{full}}$  = variation unexplained by the full model
- $SSE_{\text{red}}$  = variation unexplained by the reduced model

Intuitively, if \_\_\_\_\_ is much smaller than \_\_\_\_\_, the full model fits the data much better than the reduced model. The appropriate test statistic should depend on the difference \_\_\_\_\_ 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:  $F \geq F_{\alpha, p-k, n-p-1}$

<http://homepage.divms.uiowa.edu/~mbognar/applets/f.html>

# 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?

```
In [27]: 1 model.summary()
```

```
Out[27]:
```

OLS Regression Results

<b>Dep. Variable:</b>	sales	<b>R-squared:</b>	0.897
<b>Model:</b>	OLS	<b>Adj. R-squared:</b>	0.896
<b>Method:</b>	Least Squares	<b>F-statistic:</b>	570.3
<b>Date:</b>	Tue, 28 Nov 2017	<b>Prob (F-statistic):</b>	1.58e-96
<b>Time:</b>	20:28:02	<b>Log-Likelihood:</b>	-386.18
<b>No. Observations:</b>	200	<b>AIC:</b>	780.4
<b>Df Residuals:</b>	196	<b>BIC:</b>	793.6
<b>Df Model:</b>	3		
<b>Covariance Type:</b>	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
<b>const</b>	2.9389	0.312	9.422	0.000	2.324	3.554
<b>tv</b>	0.0458	0.001	32.809	0.000	0.043	0.049
<b>radio</b>	0.1885	0.009	21.893	0.000	0.172	0.206
<b>news</b>	-0.0010	0.006	-0.177	0.860	-0.013	0.011

# The road to $R^2$ for MLR

- Just as with simple regression, the error sum of squares is:
- 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  $p+1$   $\beta$  coefficients.

# The road to $R^2$

- Just as before, the **total sum of squares** is:
- And the **sum of squared errors** is:
- Then the coefficient of multiple determination  $R^2$  is:
- 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 $R^2$

- 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  $R^2$  to take account of the size of the model:

# Adjusted R<sup>2</sup>

- 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 R<sup>2</sup> 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 R<sup>2</sup>

```
In [27]: 1 model.summary()
```

Out[27]:

OLS Regression Results

<b>Dep. Variable:</b>	sales	<b>R-squared:</b>	0.897
<b>Model:</b>	OLS	<b>Adj. R-squared:</b>	0.896
<b>Method:</b>	Least Squares	<b>F-statistic:</b>	570.3
<b>Date:</b>	Tue, 28 Nov 2017	<b>Prob (F-statistic):</b>	1.58e-96
<b>Time:</b>	20:28:02	<b>Log-Likelihood:</b>	-386.18
<b>No. Observations:</b>	200	<b>AIC:</b>	780.4
<b>Df Residuals:</b>	196	<b>BIC:</b>	793.6
<b>Df Model:</b>	3		
<b>Covariance Type:</b>	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
<b>const</b>	2.9389	0.312	9.422	0.000	2.324	3.554
<b>tv</b>	0.0458	0.001	32.809	0.000	0.043	0.049
<b>radio</b>	0.1885	0.009	21.893	0.000	0.172	0.206
<b>news</b>	-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.

# 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:**
    1. fit null model with an intercept but no predictors.
    2. fit  $p$ -SLRs, 1 for each feature. Choose the one that gives the lowest SSE.
    3. fit  $p-1$  MLRs. Choose that which gives lowest SSE...
    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.
2. **Home prices example.** I have 1000 data points and 30 features. I want to learn the 10 most predictive and significant features.
3. **Home prices example.** I have 100 data points and 200 features. I want to learn the 20 most predictive features.
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.



# 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	3	5	5
1	2	3	6
2	1	4	7

# 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?

# Comparing multiple means

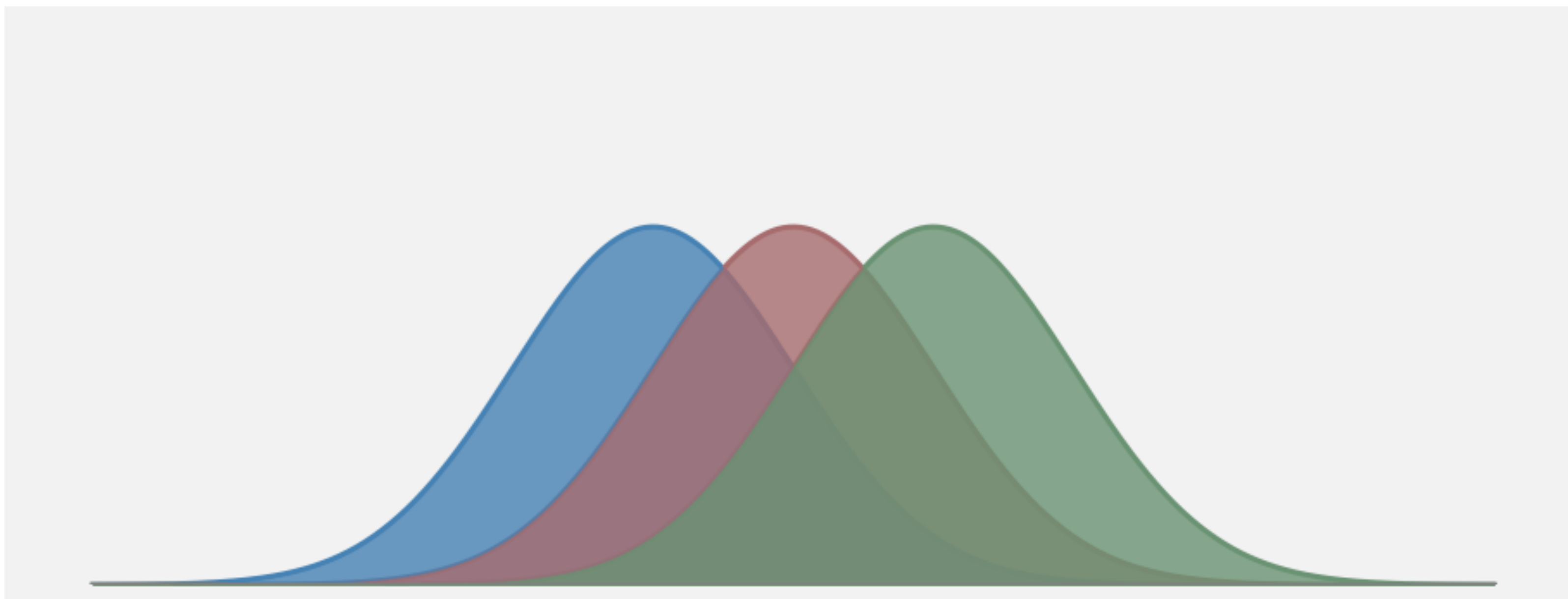
- We're often interested in comparing the means of a response from different groups
- **Example:** Suppose we are doing a study on the effect of diet on weight-loss. We have three different groups in the study:
  - **Control group:** exercise only
  - **Treatment A:** exercise plus Diet A
  - **Treatment 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?

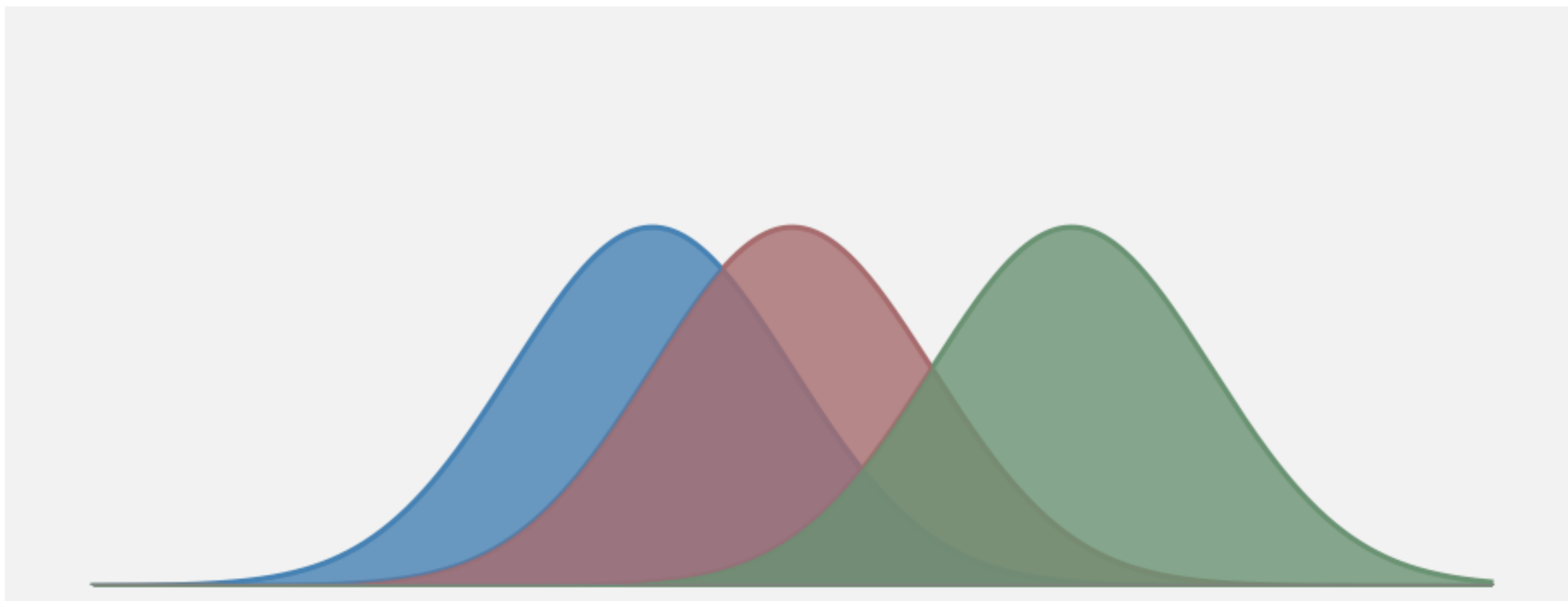
# Analysis of variance

- 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.



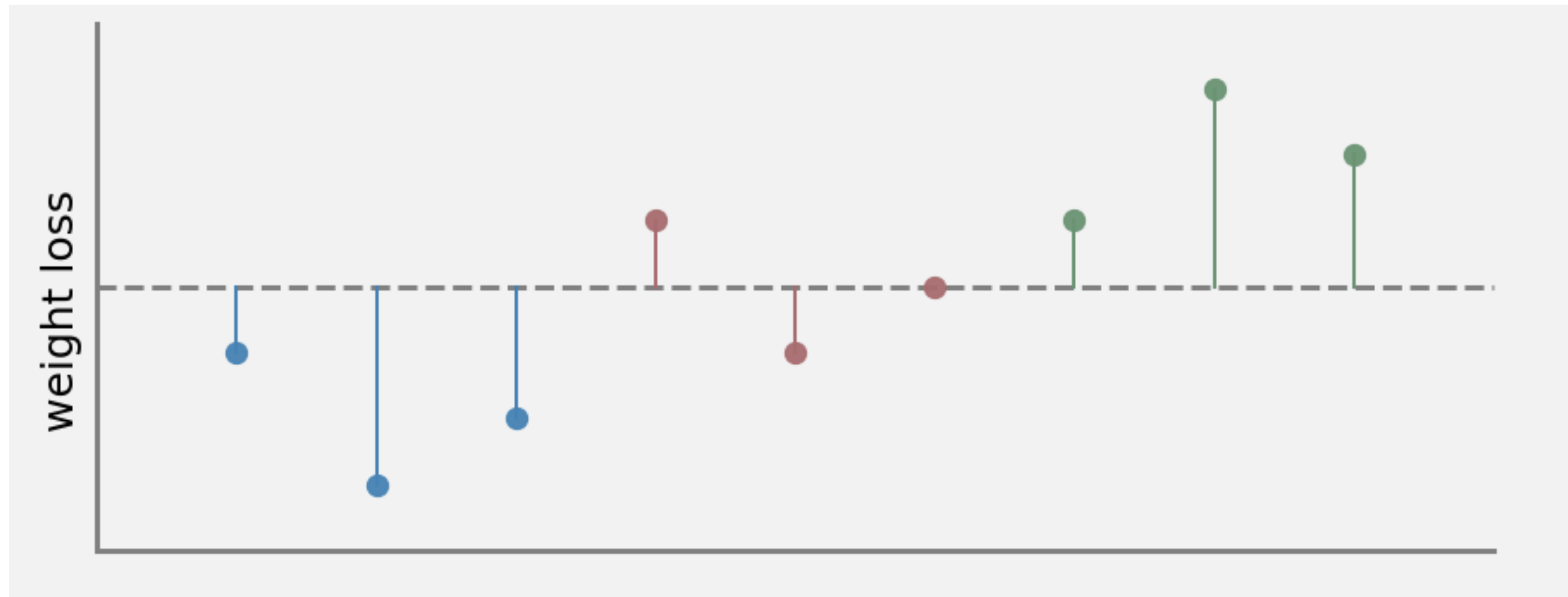
# Analysis of variance

- 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.



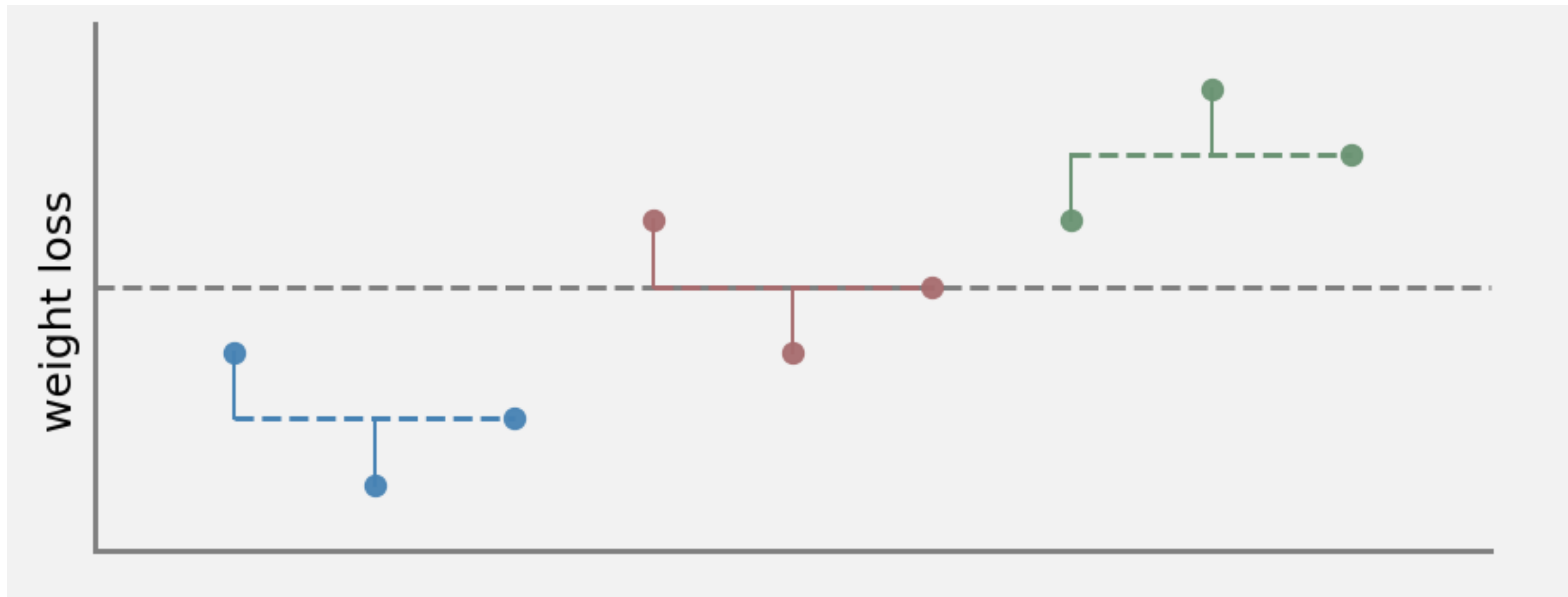
# Analysis of variance

- 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.



# 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:

## **Assumptions:**

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



# The one-way ANOVA model

Let's compute some means!

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

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

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

# It's the *variances*, stupid

- Where does the total variation in the data come from? Remember linear regression:
- A helpful decomposition:
- Then, a minor (mathematical) miracle occurs:

# The one-way ANOVA model

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

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

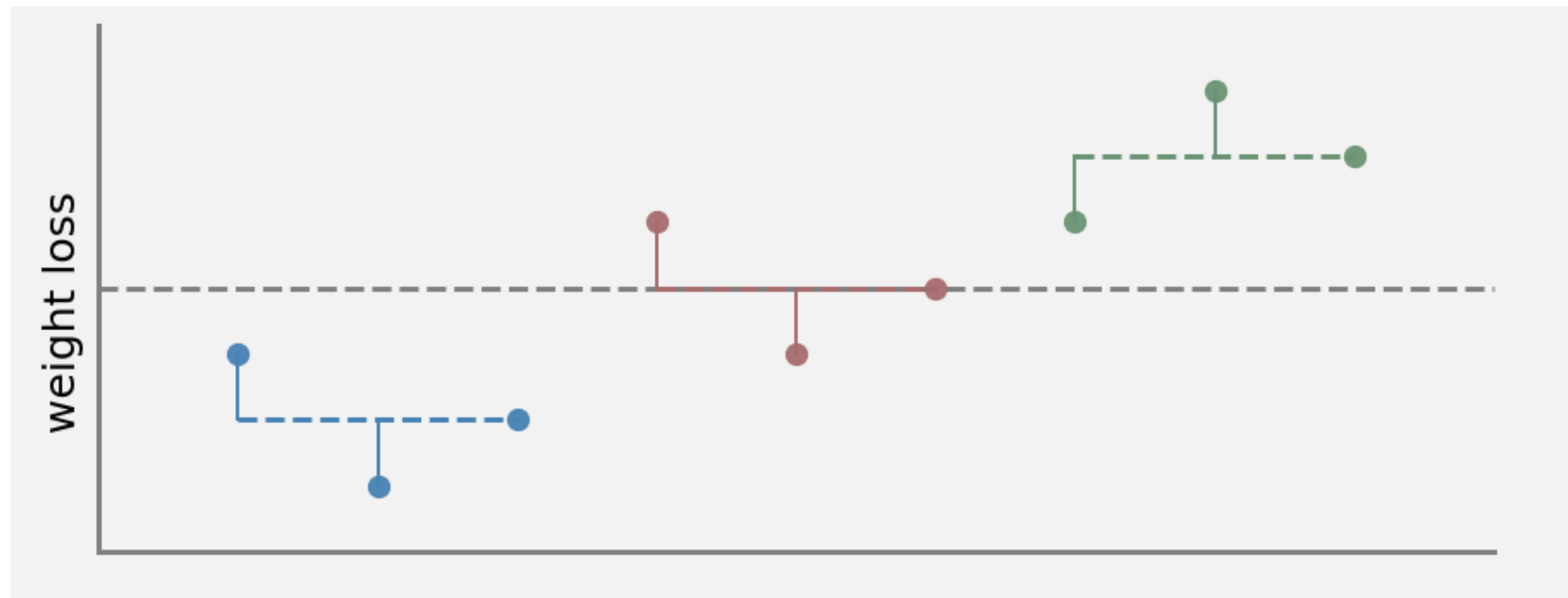
	Control	Diet A	Diet B
0	3	5	5
1	2	3	6
2	1	4	7

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

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

# The one-way ANOVA model

- Compare these results to the original picture:



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

# The one-way ANOVA model

What about degrees of freedom?

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

	Control	Diet A	Diet B
<b>0</b>	3	5	5
<b>1</b>	2	3	6
<b>2</b>	1	4	7

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

# A hypothesis test

- 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:

# The ANOVA Table

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

	Control	Diet A	Diet B
<b>0</b>	3	5	5
<b>1</b>	2	3	6
<b>2</b>	1	4	7

# ANOVA as multiple linear regression

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



# ANOVA as multiple linear regression

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

# Tukey's honest significance test

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