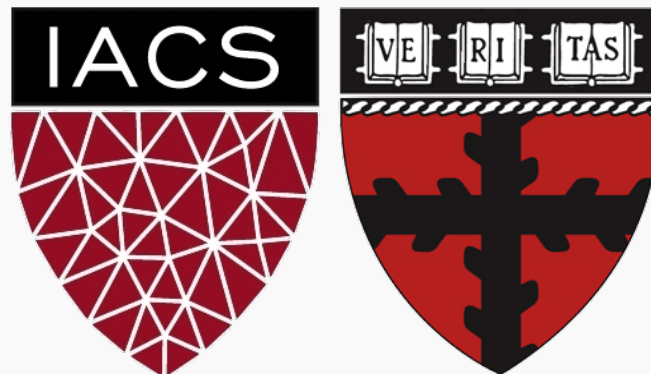


Advanced Section #1: Linear Algebra and Hypothesis Testing

CS109A Introduction to Data Science
Pavlos Protopapas, Kevin Rader and Chris Tanner



Advanced Section 1

WARNING

This deck uses animations to focus attention and break apart complex concepts.

Either watch the section video or read the deck in Slide Show mode.

Advanced Section 1

Today's topics:

Linear Algebra (Math 21b, 8 weeks)

Maximum Likelihood Estimation (Stat 111/211, 4 weeks)

Hypothesis Testing (Stat 111/211, 4 weeks)

Our time limit: 75 minutes

- We will move fast
- You are only expected to catch the big ideas
- Much of the deck is intended as notes
- We will recap the big ideas at the end of each section

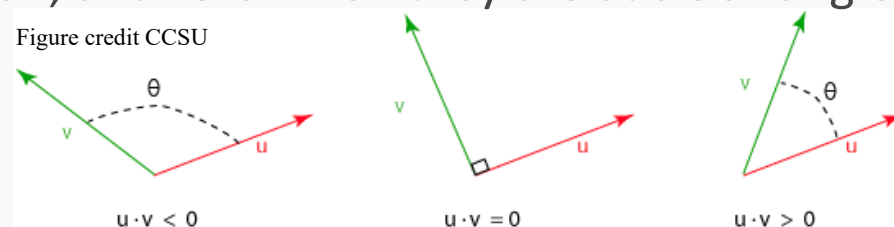
LINEAR ALGEBRA (THE HIGHLIGHTS)

Interpreting the dot product

What does a dot product mean?

$$(1,5,2) \cdot (3,-2,4) = 1 \cdot (3) + 5 \cdot (-2) + 2 \cdot (4)$$

- **Weighted sum:** We weight the entries of one vector by the entries of the other
 - Either vector can be seen as weights
 - Pick whichever is more convenient in your context
- **Measure of Length:** A vector dotted with itself gives the squared distance from (0,0,0) to the given point
 - $(1,5,2) \cdot (1,5,2) = 1 \cdot (1) + 5 \cdot (5) + 2 \cdot (2) = (1 - 0)^2 + (5 - 0)^2 + (2 - 0)^2 = 28$
 - $(1,5,2)$ thus has length $\sqrt{28}$
- **Measure of orthogonality:** For vectors of fixed length, $a \cdot b$ is biggest when a and b point are in the same direction, and zero when they are at a 90° angle



Question: how could we get a true measure of orthogonality (one that ignores length?)

$$a \cdot b / (||a|| ||b||) = a \cdot b / (\sqrt{a \cdot a} \sqrt{b \cdot b})$$

Product for Matrices

2	-1	3
1	5	2
-1	1	3
6	4	9
2	2	1

5 by 3

3	1
-2	7
4	-2

3 by 2

$$\sum_i x_{j,i} \cdot y_{i,k} = z_{j,k}$$

=

20	-11
1	32
7	0
46	16
6	14

5 by 2

$(1,5,2) \cdot (3,-2,4)$

$(2,2,1) \cdot (1,7,-2)$

Matrix multiplication is a bunch of dot products

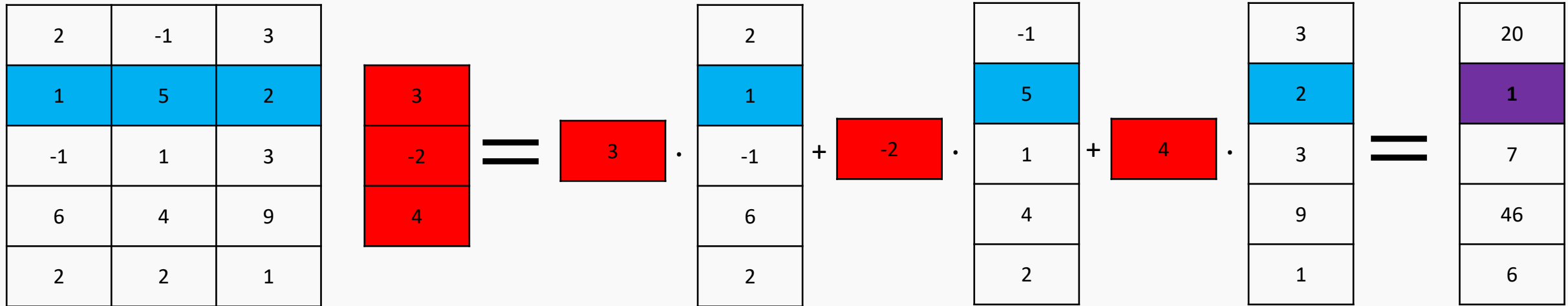
- In fact, it is every possible dot product, nicely organized
- Matrices being multiplied must have the shapes $(n, m) \times (m, p)$ and the result is of size (n, p)
 - (the middle dimensions have to match, and then drop out)

Column by Column

2	-1	3
1	5	2
-1	1	3
6	4	9
2	2	1

3	1
-2	7
4	-2

20	-11
1	32
7	0
46	16
6	14



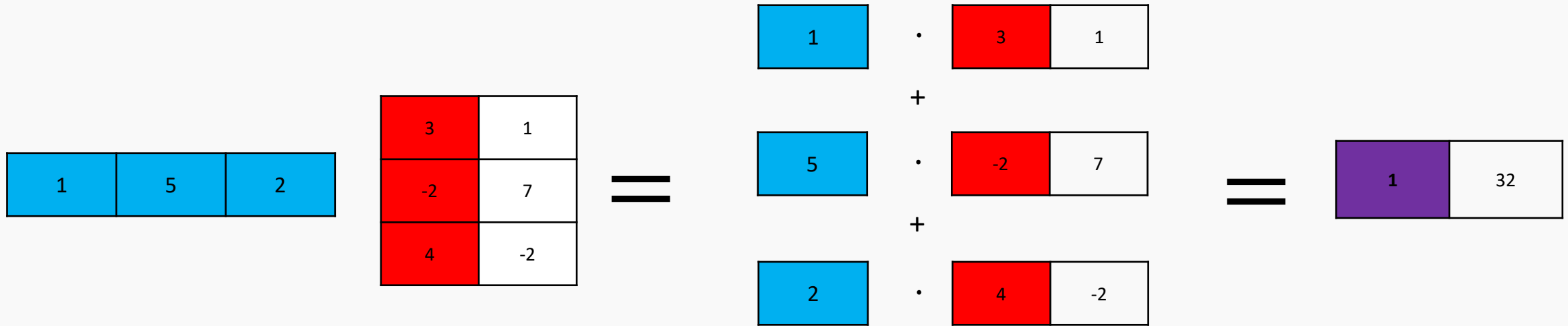
- Since matrix multiplication is a dot product, *we can think of it as a weighted sum*
 - We weight each column as specified, and sum them together
 - This produces the first column of the output
 - The second column of the output combines the same columns under different weights
- Rows?

Row by Row

2	-1	3
1	5	2
-1	1	3
6	4	0
2	2	1

3	1
-2	7
4	-2

20	-11
1	32
7	0
46	16
6	14



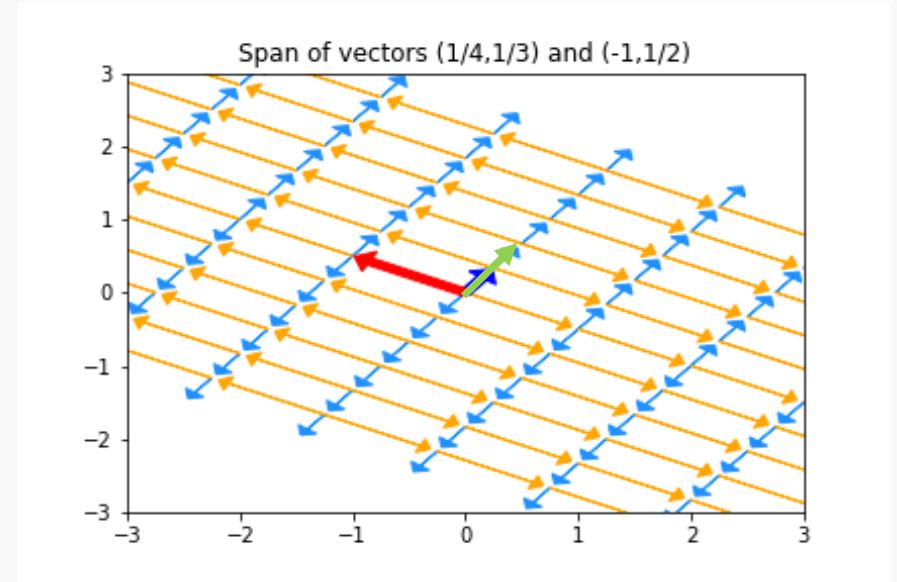
- Apply a row of A as weights on the rows of B to get a row of output

Span

LINEAR ALGEBRA (THE HIGHLIGHTS)

Span and Column Space

$$\beta_1 \cdot \begin{bmatrix} 2 \\ 1 \\ -1 \\ 6 \\ 2 \end{bmatrix} + \beta_2 \cdot \begin{bmatrix} -1 \\ 4 \\ 1 \\ 4 \\ 2 \end{bmatrix} + \beta_3 \cdot \begin{bmatrix} 3 \\ 2 \\ 3 \\ 9 \\ 1 \end{bmatrix}$$

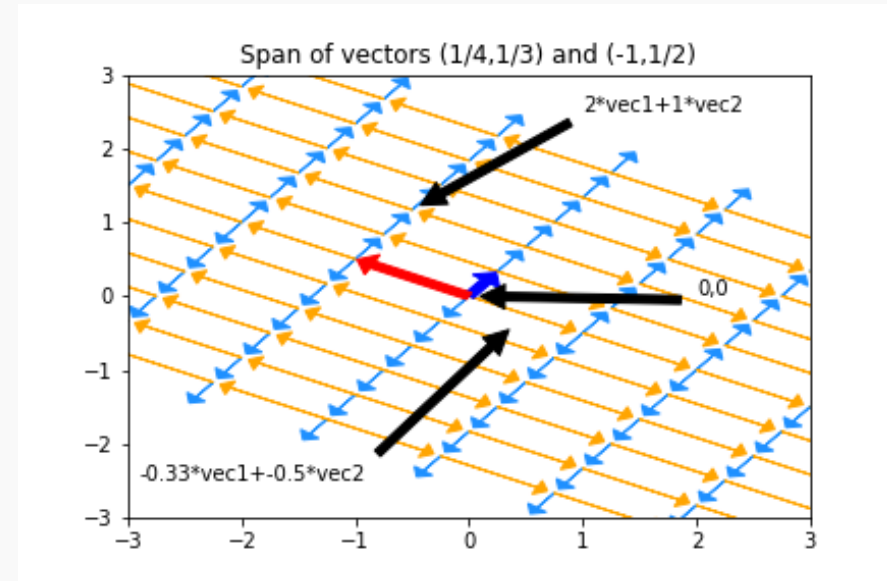
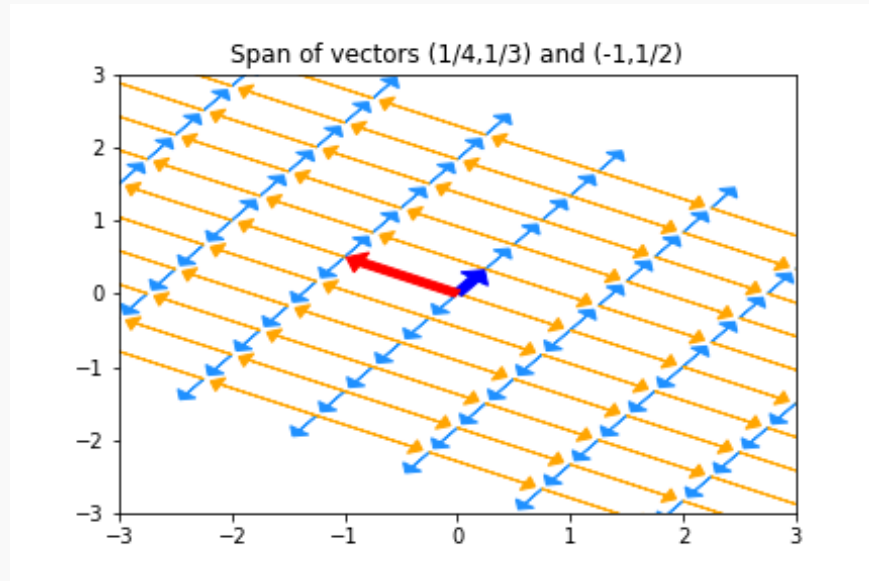


- **Span:** every possible linear combination of some vectors
 - If vectors are the columns of a matrix we call it the **column space** of that matrix
 - If vectors are the rows of a matrix it is the **row space** of that matrix
- Q: what is the span of $\{(-2,3), (5,1)\}$? what is the span of $\{(1/4,1/3), (1/2,2/3)\}$? What is the span of $\{(1,2,3), (-2,-4,-6), (1,1,1)\}$

Bases

LINEAR ALGEBRA (THE HIGHLIGHTS)

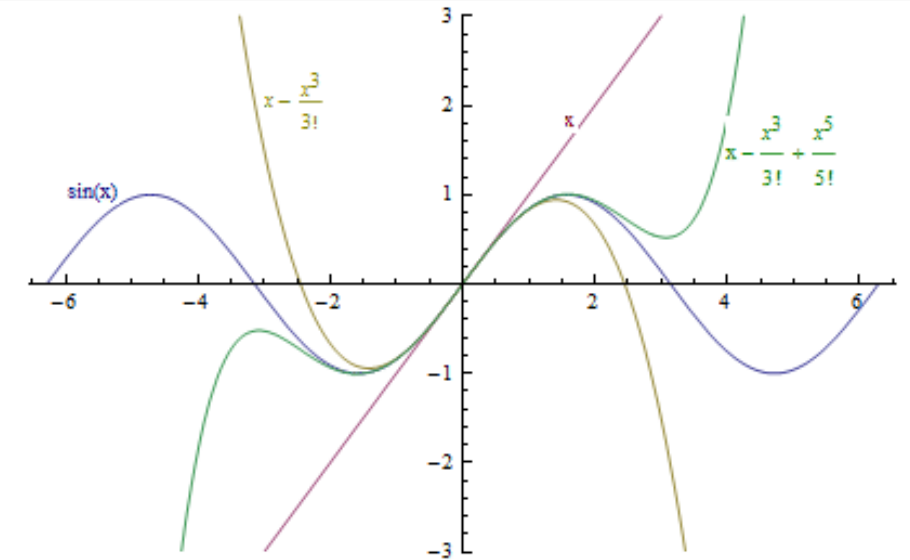
Basis Basics



- Given a space, we'll often want to come up with a set of vectors that span it
- If we give a minimal set of vectors, we've found a **basis** for that space
- A basis is a coordinate system for a space
 - Any element in the space is a weighted sum of the basis elements
 - Each element has exactly one representation in the basis
- The same space can be viewed in any number of bases - pick a good one

Function Bases

- Bases can be quite abstract:
 - Taylor polynomials express any analytic function in the infinite basis $(1, x, x^2, x^3, \dots)$
 - The Fourier transform expresses many functions in a basis built on sines and cosines
 - Radial Basis Functions express functions in yet another basis
- In all cases, we get an 'address' for a particular function
 - In the Taylor basis, $\sin(x) = (0, 1, 0, \frac{1}{6}, 0, \frac{1}{120}, \dots)$
- Bases become super important in feature engineering
 - y may depend on some transformation of x , but we only have x itself
 - We can include features $(1, x, x^2, x^3, \dots)$ to approximate

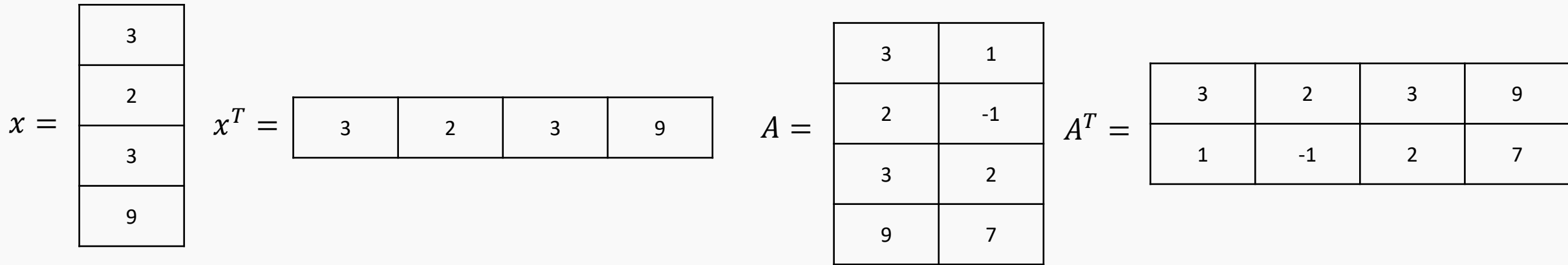


Taylor approximations to $y=\sin(x)$

Interpreting Transpose and Inverse

LINEAR ALGEBRA (THE HIGHLIGHTS)

Transpose



- Transposes switch columns and rows. Written A^T
- Better dot product notation: $a \cdot b$ is often expressed as $a^T b$
- Interpreting: The matrix multiplication AB is rows of A dotted with columns of B
 - $A^T B$ is *columns* of A dotted with columns of B
 - AB^T is rows of A dotted with rows of B
- Transposes (sort of) distribute over multiplication and addition:

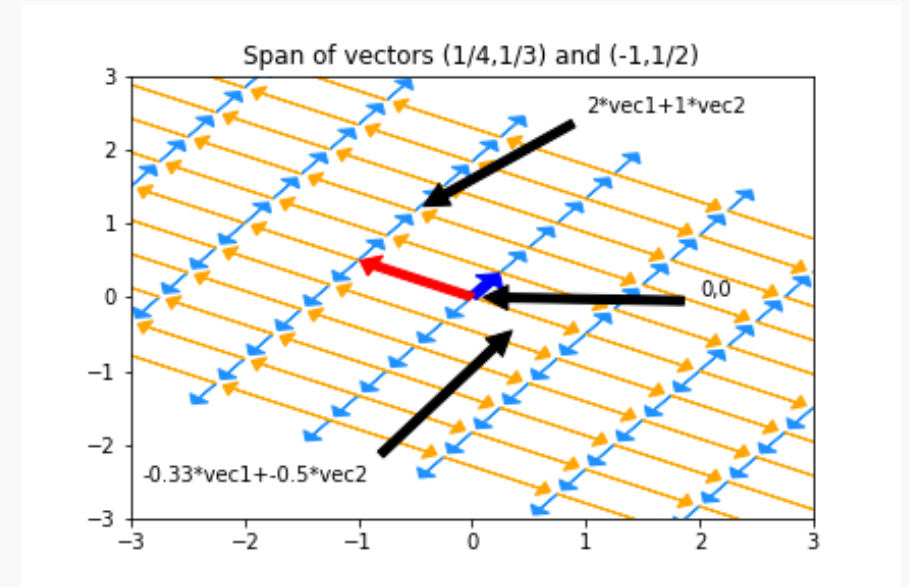
$$(AB)^T = B^T A^T$$

$$(A + B)^T = A^T + B^T$$

$$(A^T)^T = A$$

Inverses

- Algebraically, $AA^{-1} = A^{-1}A = 1$
- Geometrically, A^{-1} writes an arbitrary point b in the coordinate system provided by the columns of A
 - Proof (read this later):
 - Consider $Ax = b$. We're trying to find weights x that combine A 's columns to make b
 - Solution $x = A^{-1}b$ means that when A^{-1} multiplies a vector we get that vector's coordinates in A 's basis
- Matrix inverses exist iff columns of the matrix form a basis
 - 1 Million other equivalents to invertibility:
[Invertible Matrix Theorem](#)



How do we write $(-2,1)$ in this basis?

Just multiply A^{-1} by $(-2,1)$

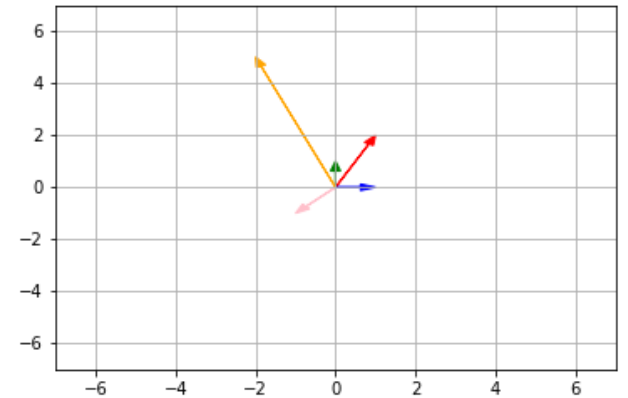
Eigenvalues and Eigenvectors

LINEAR ALGEBRA (THE HIGHLIGHTS)

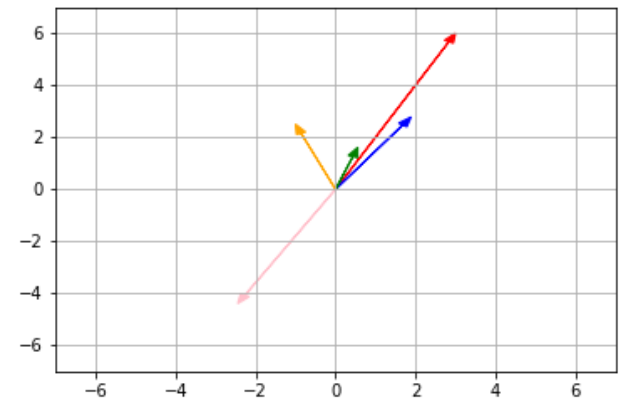
Eigenvalues

- Sometimes, multiplying a vector by a matrix just scales the vector
 - The red vector's length triples
 - The orange vector's length halves
 - All other vectors point in new directions
- The vectors that simply stretch are called *eigenvectors*. The amount they stretch is their *eigenvalue*
 - Anything along the given axis is an eigenvector; Here, $(-2,5)$ is an eigenvector so $(-4,10)$ is too
 - We often pick the version with length 1
- When they exist, eigenvectors/eigenvalues can be used to understand what a matrix does

Original vectors:



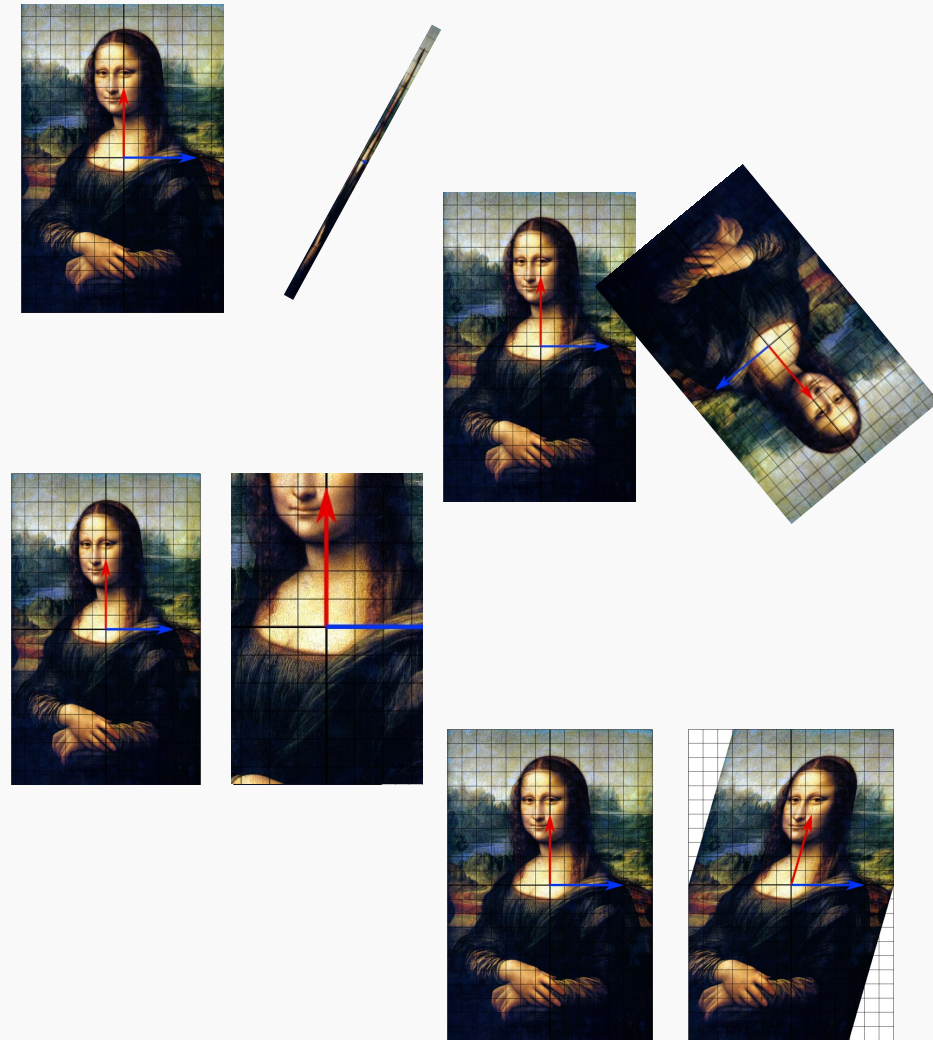
After multiplying by 2x2 matrix A:



Interpreting Eigenthings

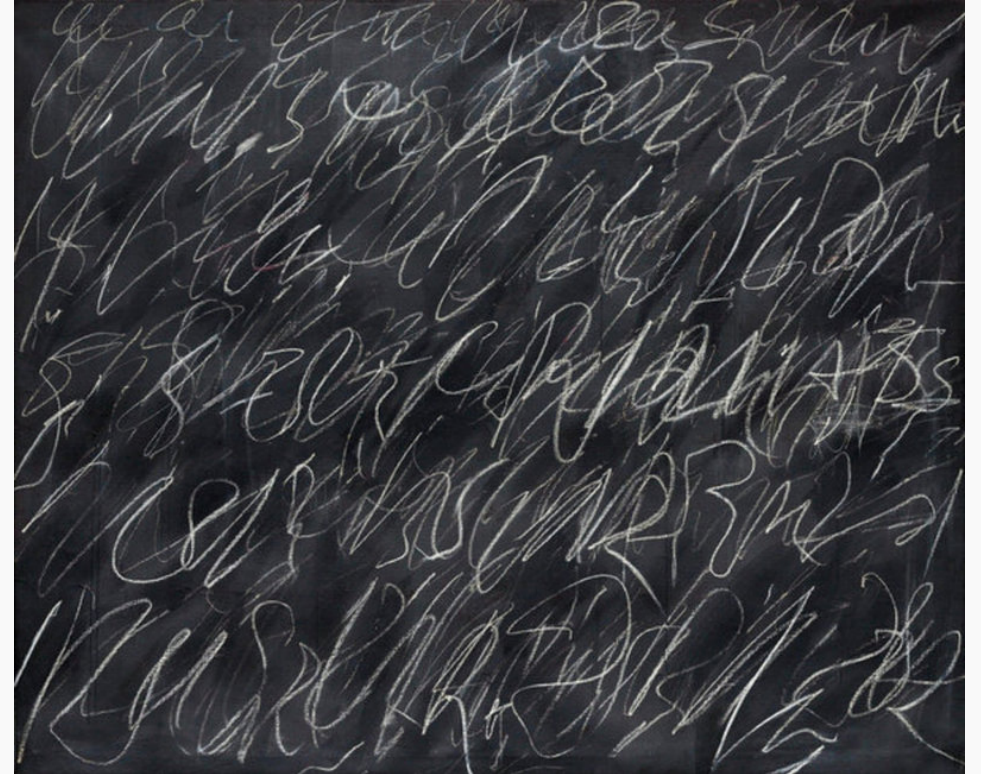
Warnings and Examples:

- Eigenvalues/Eigenvectors only apply to square matrices
- Eigenvalues may be 0 (indicating some axis is removed entirely)
- Eigenvalues may be complex numbers (indicating the matrix applies a rotation)
- Eigenvalues may be repeat, with one eigenvector per repetition (the matrix may scales some n-dimension subspace)
- Eigenvalues may repeat, with some eigenvectors missing (shears)
- If we have a full set of eigenvectors, we know everything about the given matrix S , and $S = QDQ^{-1}$
 - Q 's columns are eigenvectors, D is diagonal matrix of eigenvalues
- Question: how can we interpret this equation?



Calculating Eigenvalues

- Eigenvalues can be found by:
 - **A computer program**
- But what if we need to do it on a blackboard?
 - The definition $Ax = \lambda x$
 - This says that for special vectors x , multiplying by the matrix A is the same as just scaling by λ (x is then an eigenvector matching eigenvalue λ)
 - The equation $\det(A - \lambda I_n) = 0$
 - I_n is the n by n identity matrix of size n by n . In effect, we subtract lambda from the diagonal of A
 - Determinants are tedious to write out, but this produces a polynomial in λ which can be solved to find eigenvalues



- Eigenvectors matching known eigenvalues can be found by solving $(A - \lambda I_n)x = 0$ for x

Matrix Decomposition

LINEAR ALGEBRA (THE HIGHLIGHTS)



Matrix Decompositions

- **Eigenvalue Decomposition:** Some square matrices can be decomposed into scalings along particular axes
 - Symbolically: $S = QDQ^{-1}$; *D diagonal matrix of eigenvalues; Q made up of eigenvectors, but possibly wild (unless S was symmetric; then Q is orthonormal)*
- **Polar Decomposition:** Every matrix M can be expressed as a rotation (which may introduce or remove dimensions) and a stretch
 - Symbolically: $M = UP$ or $M=PU$; *P positive semi-definite, U's columns orthonormal*
- **Singular Value Decomposition:** Every matrix M can be decomposed into a rotation in the original space, a scaling, and a rotation in the final space
 - Symbolically: $M = U\Sigma V^T$; *U and V orthonormal, Σ diagonal (though not square)*

Where we've been

Vector dot product and Matrix product

2	-1	3
1	5	2
-1	1	3
6	4	9
2	2	1

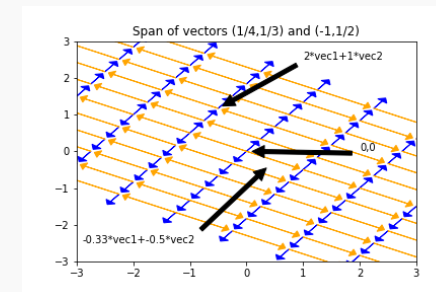
3	1
1	7
6	-2

20	-11
1	32
7	0
46	16
6	14

Span

$$\beta_1 \cdot \begin{bmatrix} 2 \\ 1 \\ -1 \\ 6 \\ 2 \end{bmatrix} + \beta_2 \cdot \begin{bmatrix} -1 \\ 4 \\ 1 \\ 4 \\ 2 \end{bmatrix} + \beta_3 \cdot \begin{bmatrix} 3 \\ 2 \\ 3 \\ 9 \\ 1 \end{bmatrix}$$

Basis as a coordinate system for a space



Other decompositions

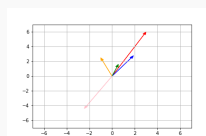
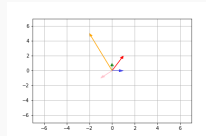
$$M = UP \text{ or } M=PU$$

$$M = U\Sigma V^T$$

Eigenvalues

$$Ax = \lambda x$$

$$S = QDQ^{-1}$$



Invertibility

$$Ax = b ; x = A^{-1}b$$

Reading

- What about all the facts about inverses and dot products I've forgotten since undergrad? [[Matrix Cookbook](#)] [[Linear Algebra Formulas](#)]

LINEAR ALGEBRA (SUMMARY)

Notes

- **Matrix multiplication:** every dot product between rows of A and columns of B
 - Important special case: a matrix times a vector is a weighted sum of the matrix columns
- **Dot products** measure similarity between two vectors: 0 is extremely un-alike, bigger is pointing in the same direction and/or longer
 - Alternatively, a dot product is a weighted sum
- **Bases:** a coordinate system for some space. Everything in the space has a unique address
- **Matrix Factorization:** all matrices are rotations and stretches. We can decompose ‘rotation and stretch’ in different ways
 - Sometimes, re-writing a matrix into factors helps us with algebra
- **Matrix Inverses** don’t always exist. The ‘stretch’ part may collapse a dimension. M^{-1} can be thought of as the matrix that expresses a given point in terms of columns of M
- **Span and Row/Column Space:** every weighted sum of given vectors
- **Linear (In)Dependence** is just “can some vector in the collection be represented as a weighted sum of the others” if not, vectors are Linearly Independent

LINEAR REGRESSION

Review and Practice: Linear Regression

- In linear regression, we're trying to write our response data y as a linear function of our [augmented] features X

$$\begin{aligned} \text{response} &= \beta_1 \text{feature}_1 + \beta_2 \text{feature}_2 + \beta_3 \text{feature}_3 + \dots \\ \hat{y} &= X\beta \end{aligned}$$

- Our response isn't necessarily a linear function of our features, so we instead find betas that produce a column \hat{y} that is as close as possible to y (in Euclidean distance): $(y - \hat{y})$

$$\min_{\beta} \sqrt{(y - \hat{y})^T (y - \hat{y})} = \min_{\beta} \sqrt{(y - X\beta)^T (y - X\beta)}$$

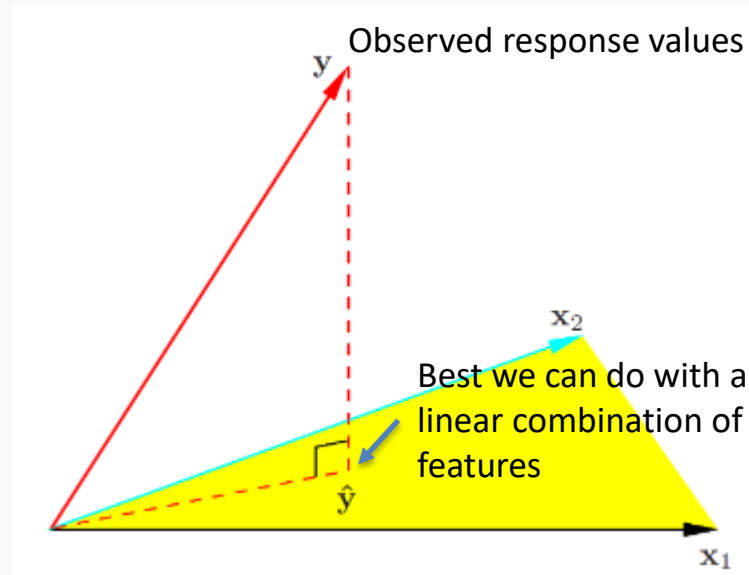
- Goal: find that the optimal $\beta = (X^T X)^{-1} X^T y$
- Steps:
 1. Drop the sqrt [why is that legal?]
 2. Distribute the transpose
 3. Distribute/FOIL all terms
 4. Take the derivative with respect to β (Matrix Cookbook (69) and (81): derivative of $\beta^T a$ is a^T , ...)
 5. Simplify and solve for beta

Interpreting LR: Algebra

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- The best possible betas, $\hat{\beta} = (X^T X)^{-1} X^T y$ can be viewed in two parts:
 - Numerator ($X^T y$): columns of X dotted with (the) column of y ; how related are the feature vectors and y ?
 - Denominator ($X^T X$): columns of X dotted with columns of X ; how related are the different features?
- Roughly, our solution assigns big values to features that predict y , but punishes features that are similar to (combinations of) other features
- Bad things happen if $X^T X$ is uninvertible (or nearly so)

Interpreting LR: Geometry



$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1}X^T y$$

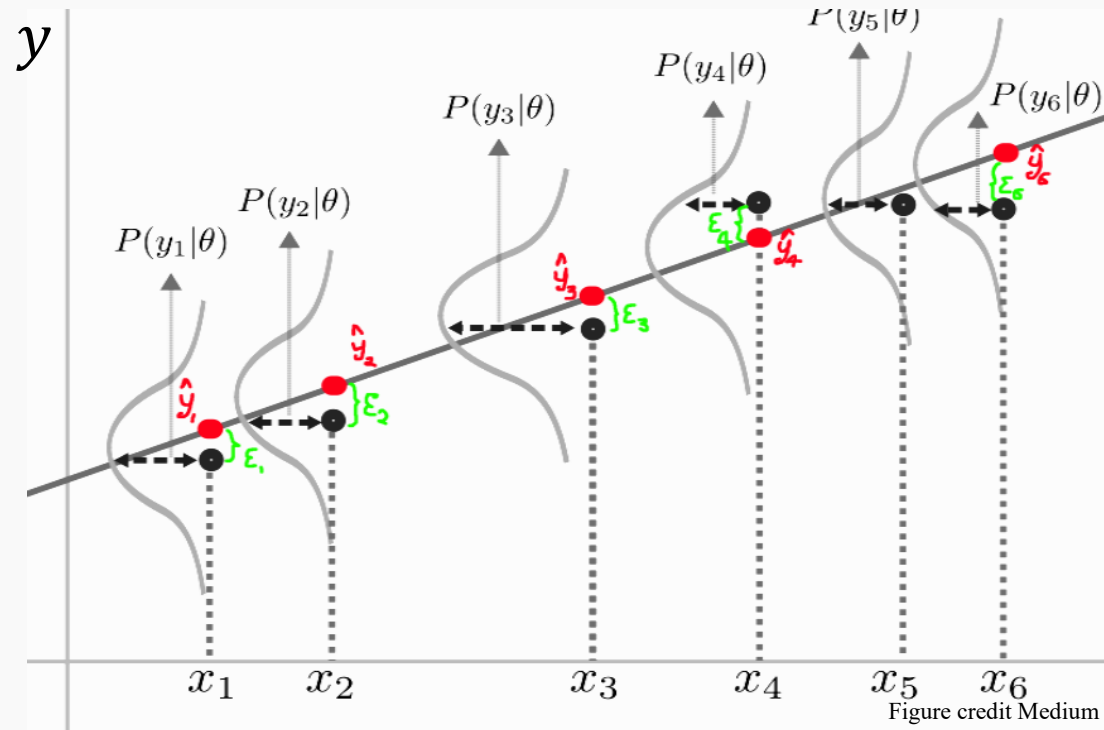
- The only points that CAN be expressed as $X\beta$ are those in the span/column space of X .
 - By minimizing distance, we're finding the point in the column space that is closest to the actual y vector
- The point $X\hat{\beta}$ is the *projection* of the observed y values onto the things linear regression can express
- Warnings:
 - Adding more columns (features) can only make the span bigger and the fit better
 - If some features are very similar, results will be unstable

Interpreting LR: MLE

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

What if we want to know more about how the data was generated? More interpretability. We would like to know not only the optimal $\hat{\beta}$ but their error bars. How good are other sets of $\hat{\beta}$.

If need to make a few assumptions:



Interpreting LR: MLE

Likelihood function:

$$P(Y=y | X, \beta, \sigma^2) = N(X\beta, \sigma^2 I_n) = \frac{1}{\sqrt{2\pi(\sigma^2 I_n)}} e^{-\frac{1}{2}(\mathbf{y}-X\beta)^T(\mathbf{y}-X\beta)/(\sigma^2 I_n)}$$

Optimal β :

$$\hat{\beta} = (X^T X)^{-1} X^T \mathbf{y}$$

Optimal σ^2 :

$$\hat{\sigma}^2 = \frac{\text{residuals under optimal } \beta}{\text{number of observations} - \text{number of features}}$$

This allows us to attach error bars to our parameter estimates



Linear Regression: Review

- LR offers a closed form solution for the optimal parameters β
- By making assumptions about where the data came from, we get richer statements from our model
- A likelihood function tells us how likely any given data set is under our model and for a set of parameters.
- MLE finds the parameters that maximize is, making our data as likely as possible
- Finding the MLE can be hard, sometimes possible via calculus, often requires computer code... depending on our assumptions.

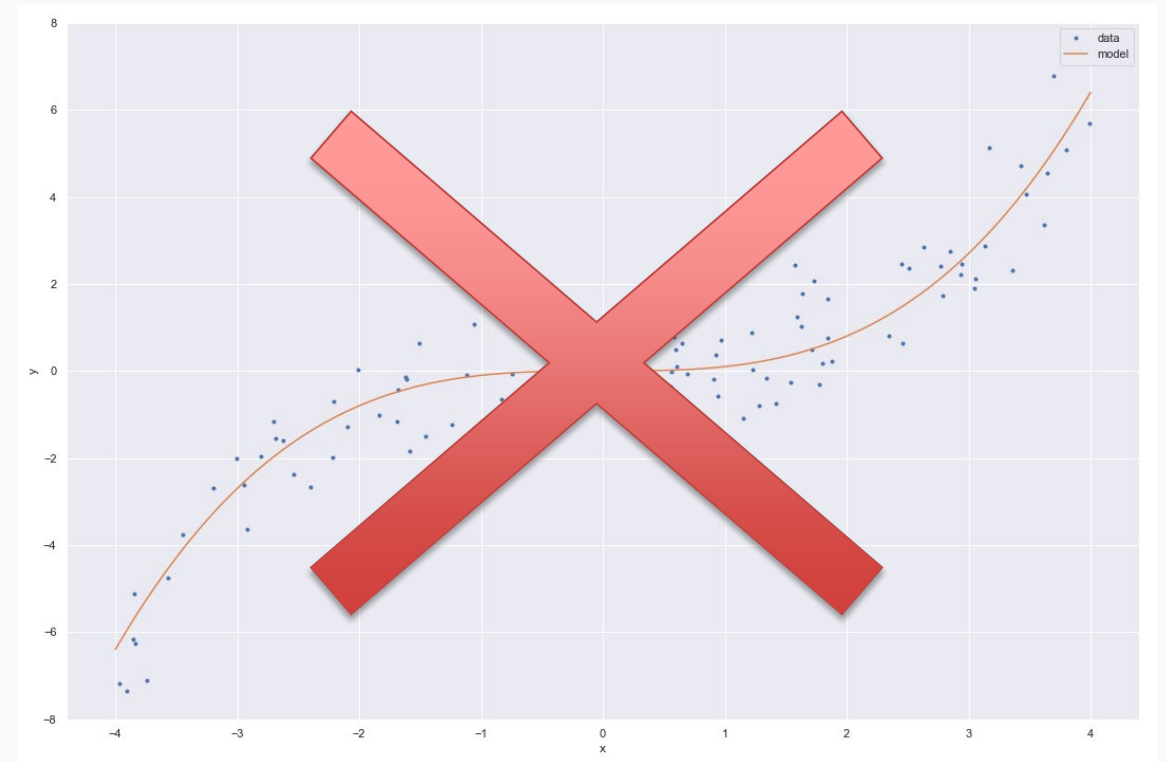


STATISTICS: HYPOTHESIS TESTING

OR: WHAT PARAMETERS EXPLAIN THE DATA

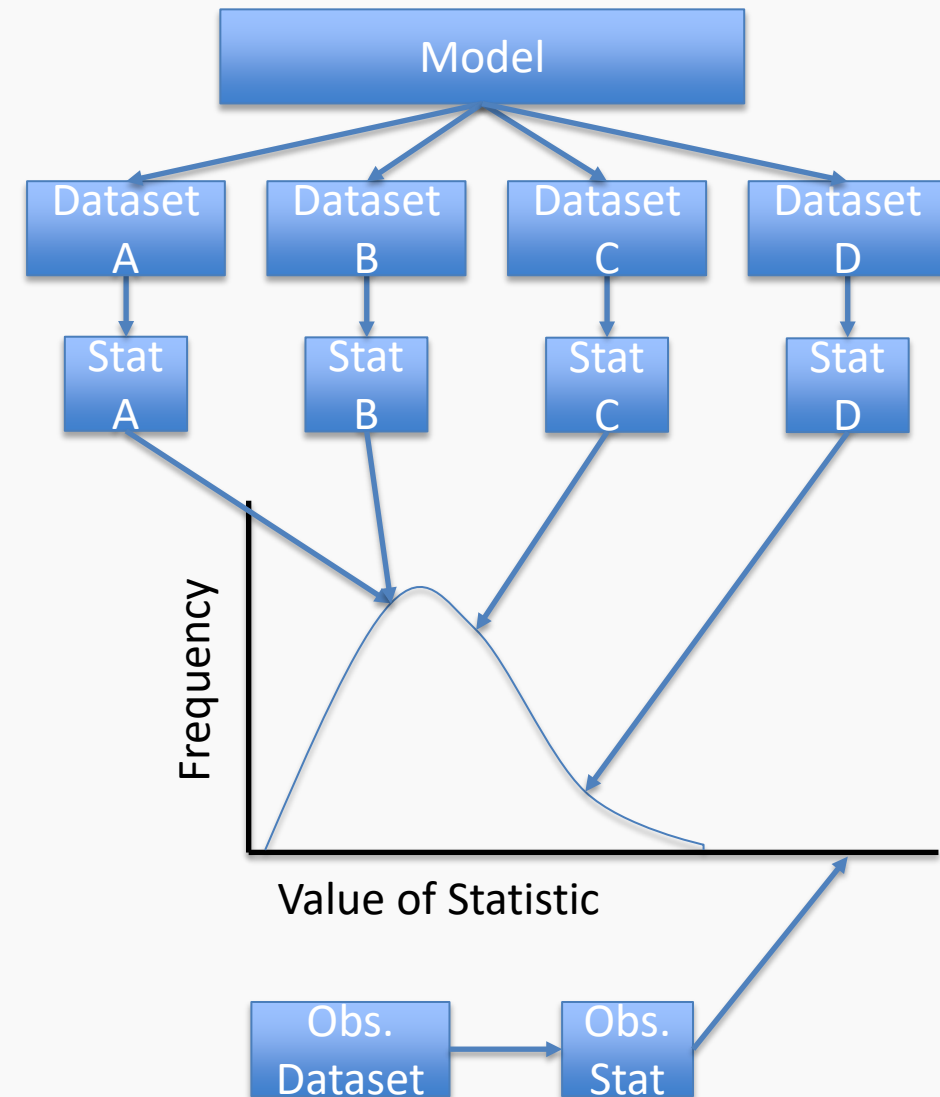
Inaccessible Truth

- We can only rule models out.
- It's impossible to prove a model is correct
 - Can you prove increasing a parameter by .0000001% is incorrect?



Model Rejection

- Important: a ‘model’ is a (probabilistic) story about how the data came to be, complete with *specified values of every parameter*.
 - The model could produce many possible datasets
 - We only have one observed dataset
- How can we tell if a model is wrong?
 - If the model is unlikely to reproduce the aspects of the data that we care about and observe, it has to go
 - Therefore, we have some real-number summary of the dataset (a ‘statistic’) by which we’ll compare model-generated datasets and our observed dataset
 - If the statistics produced by the model are clearly different than the one from the real data, we reject the model



Recap: How to understand any statistical test

- A statistical test typically specifies:
 1. A ‘hypothesized’ (probabilistic) data generating process *(Jargon: the null hypothesis)*
 2. A summary we’ll use to compress/summarize a dataset *(Jargon: a statistic)*
 3. A rule for comparing the observed and the simulated summaries

- Example: *t*-test

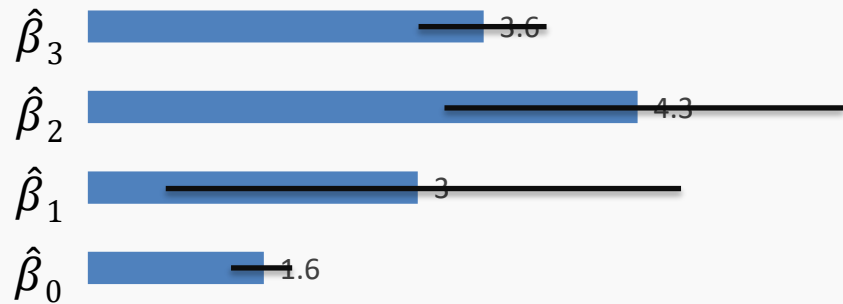
1. The *y* data are generated via the estimated line/plane, plus Normal(0, σ^2) noise, EXCEPT a particular coefficient is assumed to actually be zero!
2. The coefficient we’d calculate for that dataset (minus 0), over the SE of the coefficient

$$t \text{ statistic} = \frac{\hat{\beta}_{\text{observed}} - 0}{\widehat{SE}(\hat{\beta}_{\text{observed}})}$$

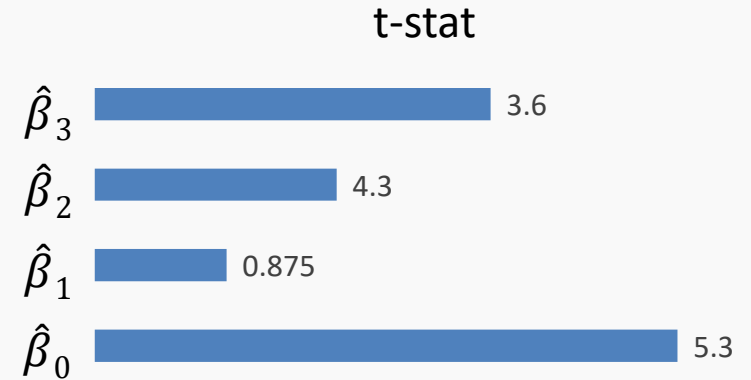
3. Declare the model bad if the observed result is in the top/bottom $\alpha/2$ of simulated results (commonly top/bottom 2.5%)

The t-test

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \varepsilon$$



$$t_{stat} = \hat{\beta} / \sigma_{\hat{\beta}}$$

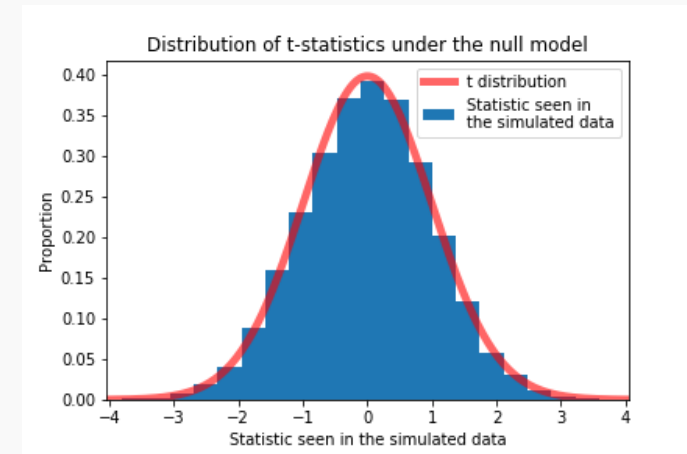
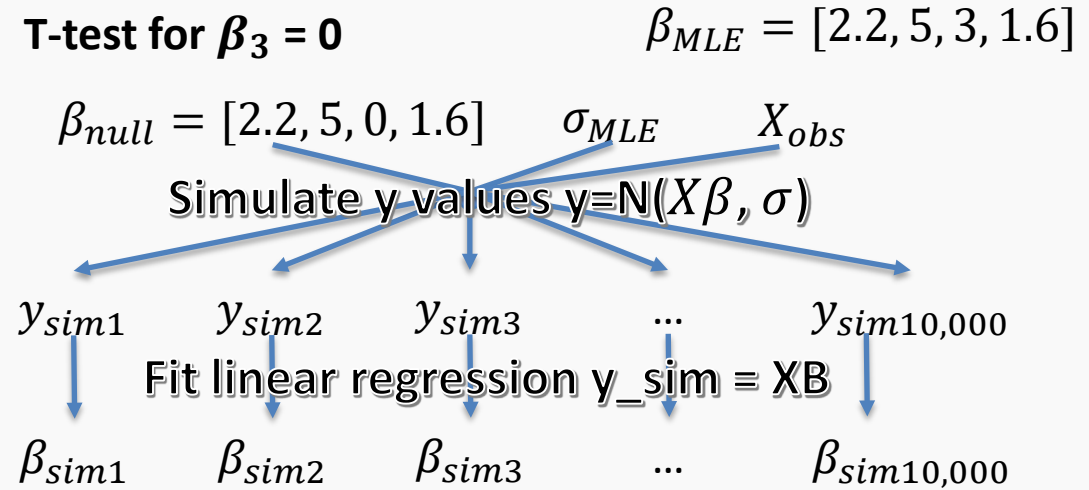


How do we interpret this? We know the relative parameter dependence, but how good is good enough?
Which of these features really matter?

The t-test

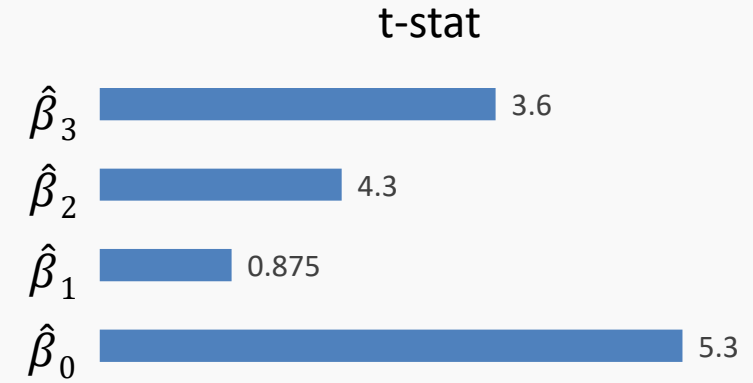
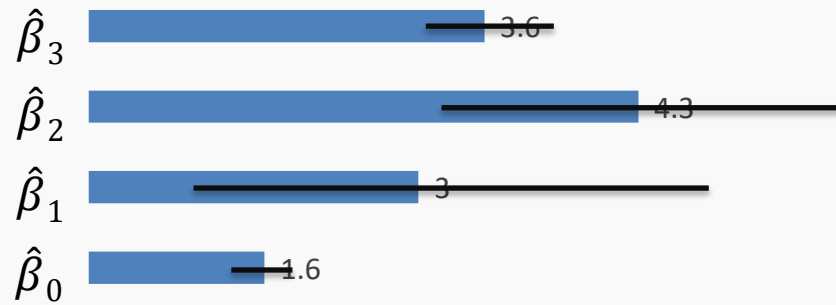
Walkthrough:

- We set a particular β (or set of β 's) we care about to zero (call them β_{null}).
- We simulate 10,000 new datasets using β_{null} as truth.
- In each of the 10,000 datasets, fit a regression against X and plot the values of the β we care about (the one we set to zero).
 - Plotting the t statistic in each simulation is a little nicer
- The t statistic calculated from the observed data was 17.8. *Do we think the proposed model generated our data?*
- One more thing: Amazingly, 'Student' knew what results we'd get from the simulation.



The t-test

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \varepsilon$$

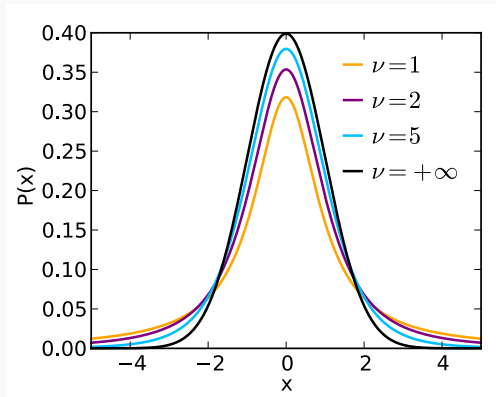


$$t_{stat} = \beta / \sigma_{\beta}$$

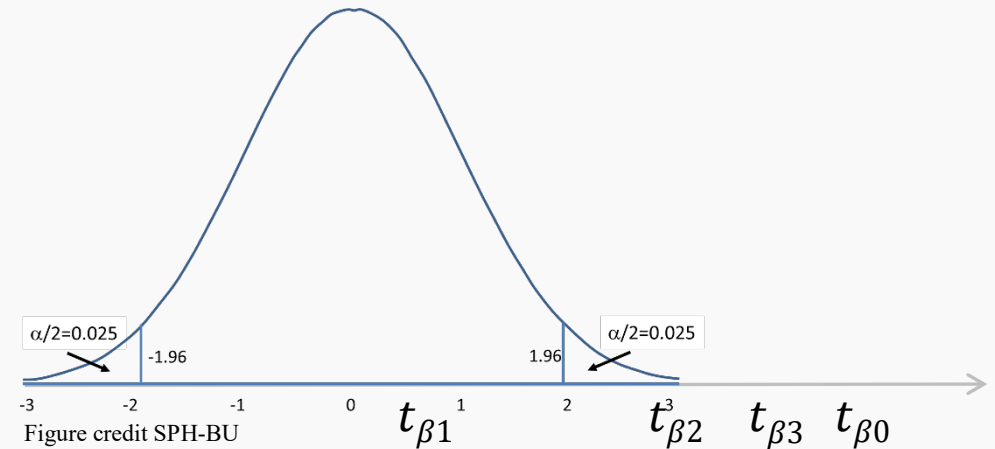
$N(0, \sigma)$

$\sigma^2 \sim \chi^2$

Student's distribution

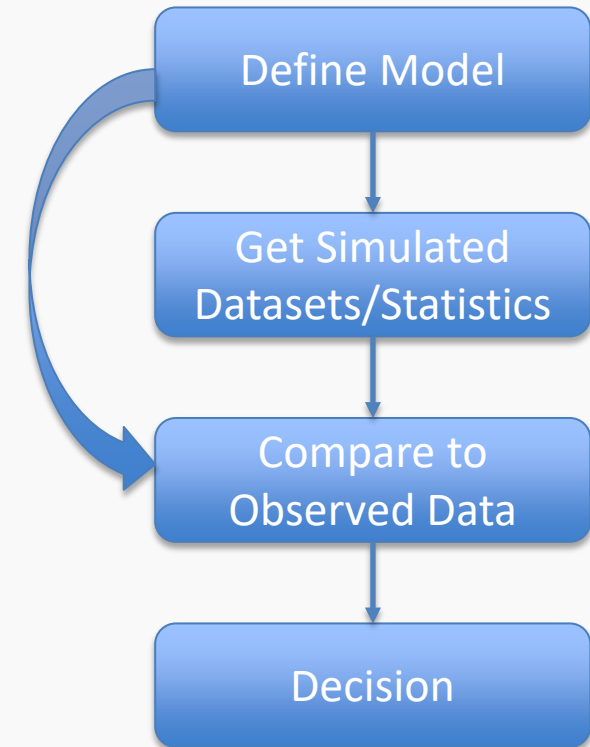


$\nu = n - 1$
 # of observations



The Value of Assumptions

- Student's clever set-up let's us skip the simulation
- In fact, all classical tests are built around working out what distribution the results will follow, without simulating
 - Student's work lets us take *infinite* samples at almost no cost
- These shortcuts were *vital* before computers, and are still important today
 - Even so, via simulation we're freer to test and reject more diverse models and use wilder summaries
 - However, the summaries and rules we choose still require thought: some are *much* better than others

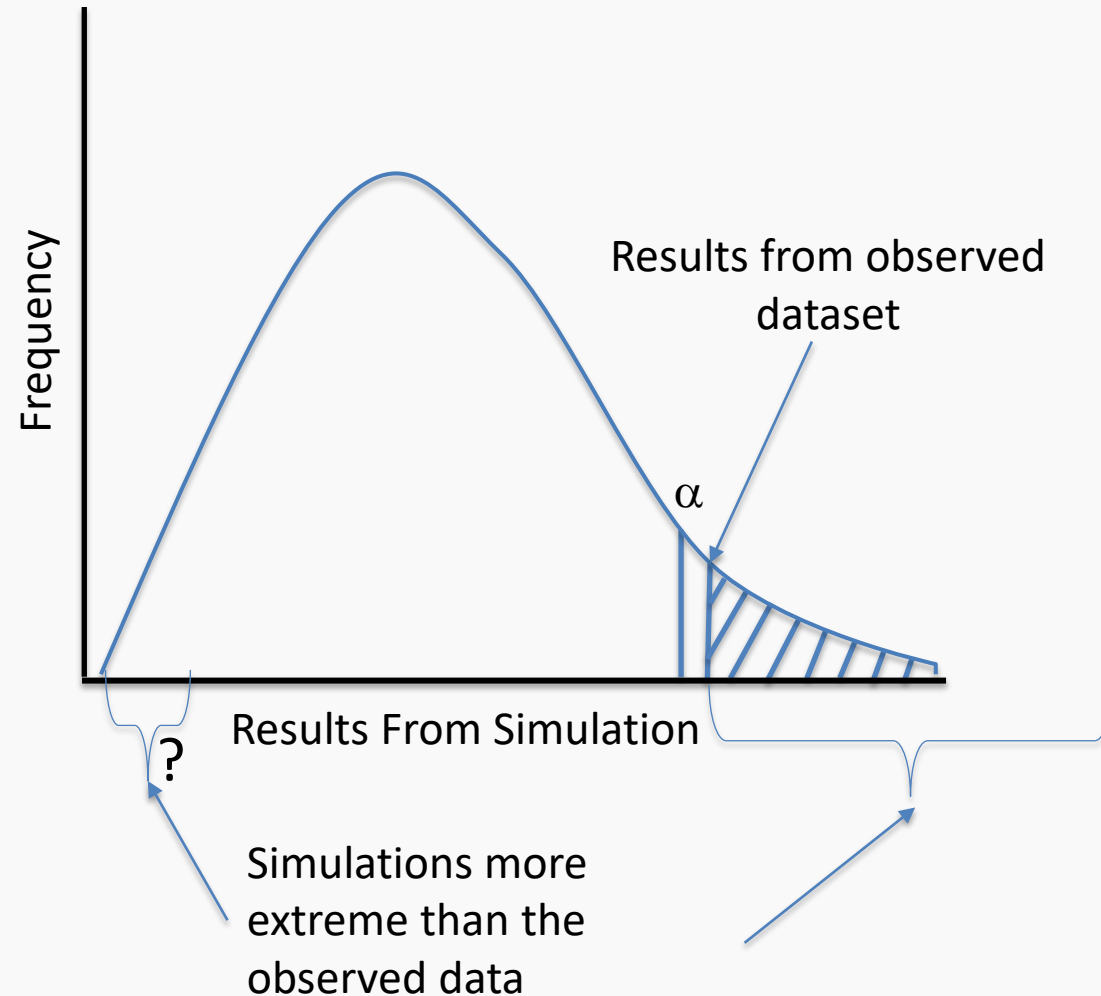


p-values

- Hypothesis (model) testing leads to comparing a distribution against a specific value
- α is the significance level: the probability to make a mistake by rejecting the null hypothesis
- A natural way to summarize: report what percentage of results are more extreme than the observed data
 - Basically, could the model frequently produce data that looks like ours?
- This is the p value: $p=0.031$ means that your observed data is in the top 3.1% of extreme results under this model (using our statistic)
 - There is some ambiguity about what 'extreme' should mean

Jargon: **p-values** are “the probability, assuming the null model is true, of seeing a value of [your statistic] as extreme or more extreme than what was seen in the observed data”

Distribution of Simulation Results



p Value Warnings

- p values are just one possible measure of the evidence against a model
- Rejecting a model when $p < \text{threshold}$ is only one possible decision rule
- **Even if the null model is exactly true, 5% of the time, we'll get a dataset with $p < .05$**
 - $p < .05$ doesn't *prove* the null model is wrong, it just suggests it.
 - It does mean that anyone who wants to believe in the null must explain with why something unlikely happened

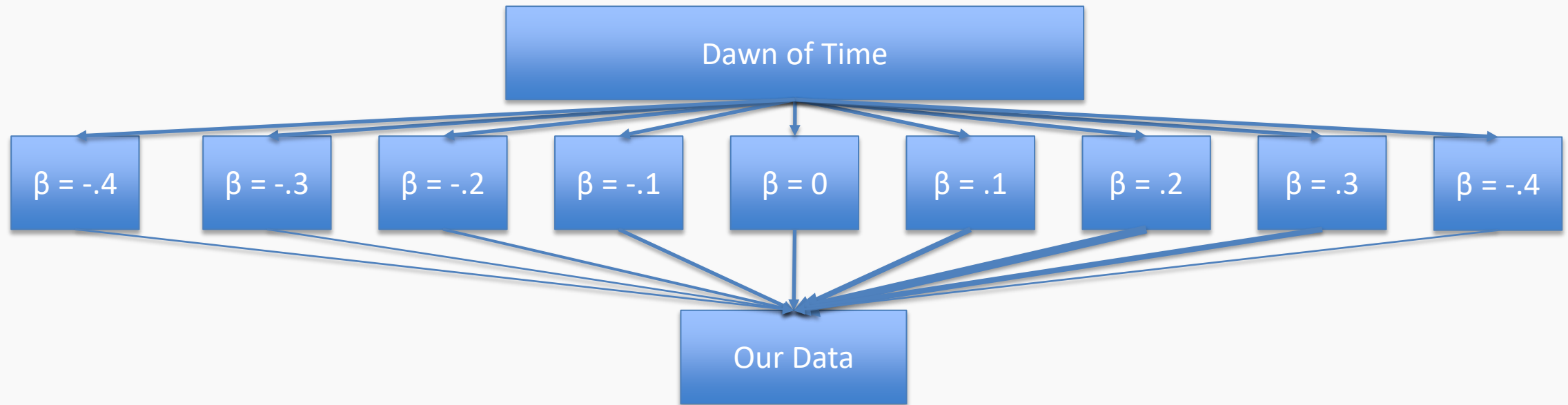
Recap

- We can't rule models in (it's difficult); we can only rule them out (much easier)
- We rule models out when the data they produce is different from the observed data
 - We pick a particular candidate (null) model
 - A statistic summarizes the simulated and observed datasets
 - We compare the statistic on the observed data to the [simulated or theoretical] *sampling distribution* of statistics the null model produces
 - We rule out the null model if the observed data doesn't seem to come from the model (disagrees with the sampling distribution).
- A p value summarizes the level of evidence against a particular null
 - "The observed data are in the top 1% of results produced by this model... what's more reasonable: we got lucky, or the model was wrong?"

STATISTICS: HYPOTHESIS TESTING

CONFIDENCE INTERVALS AND COMPOSITE HYPOTHESES

Recap

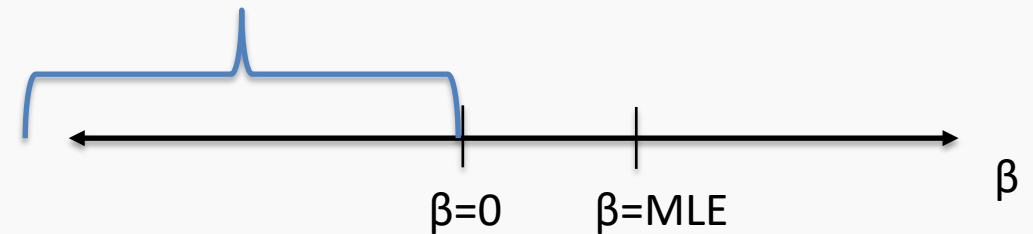


- Let's talk about what we just did
 - That t-test was ONLY testing the model where the coefficient in question is set to zero
 - Ruling out this model makes it more likely that other models are true, but doesn't tell us which ones
 - If the null is $\beta = 0$, getting $p < .05$ only rules out THAT ONE model
- When *would* it make sense to stop after ruling out $\beta = 0$, without testing $\beta = .1$?

Composite Hypotheses: Multiple Models

- Often, we're interested in trying out more than one candidate model
 - E.g. Can we disprove all models with a negative value of beta?
 - This amounts to simulating data from each of those models (but there are infinitely many...)
- Sometimes, ruling out the nearest model is enough; we know that the other models have to be worse
- If a method claims it can test $\theta < 0$, this is how

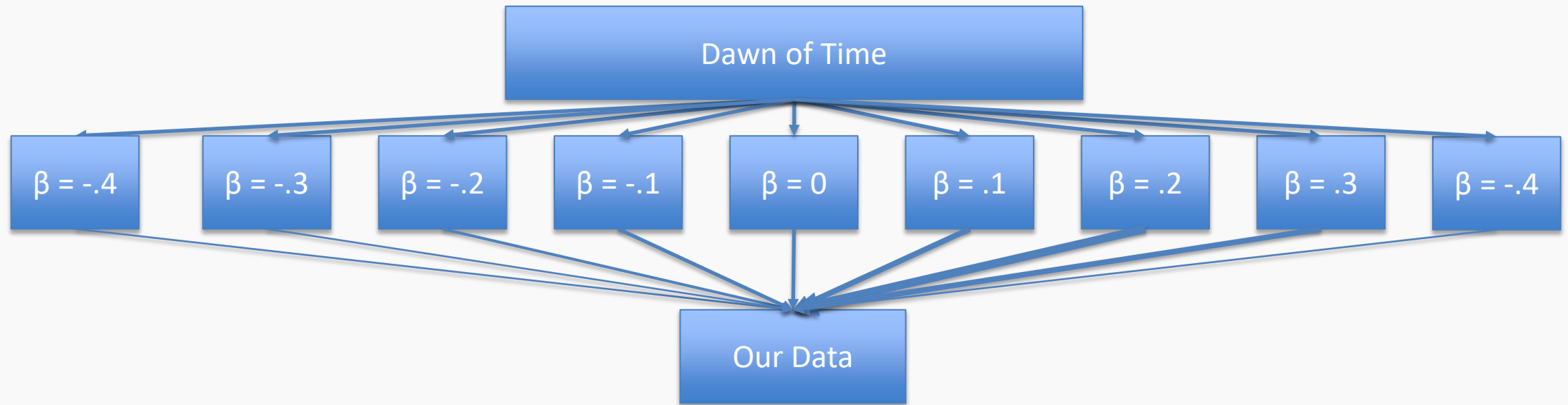
Can we rule these out?



$\beta=0$ will be closer to matching the data (in terms of t statistic) than any other model in the set*; we only need to test $\beta=0$

* Non-trivial; true for student's t but not for other measures

THE Null vs A Null

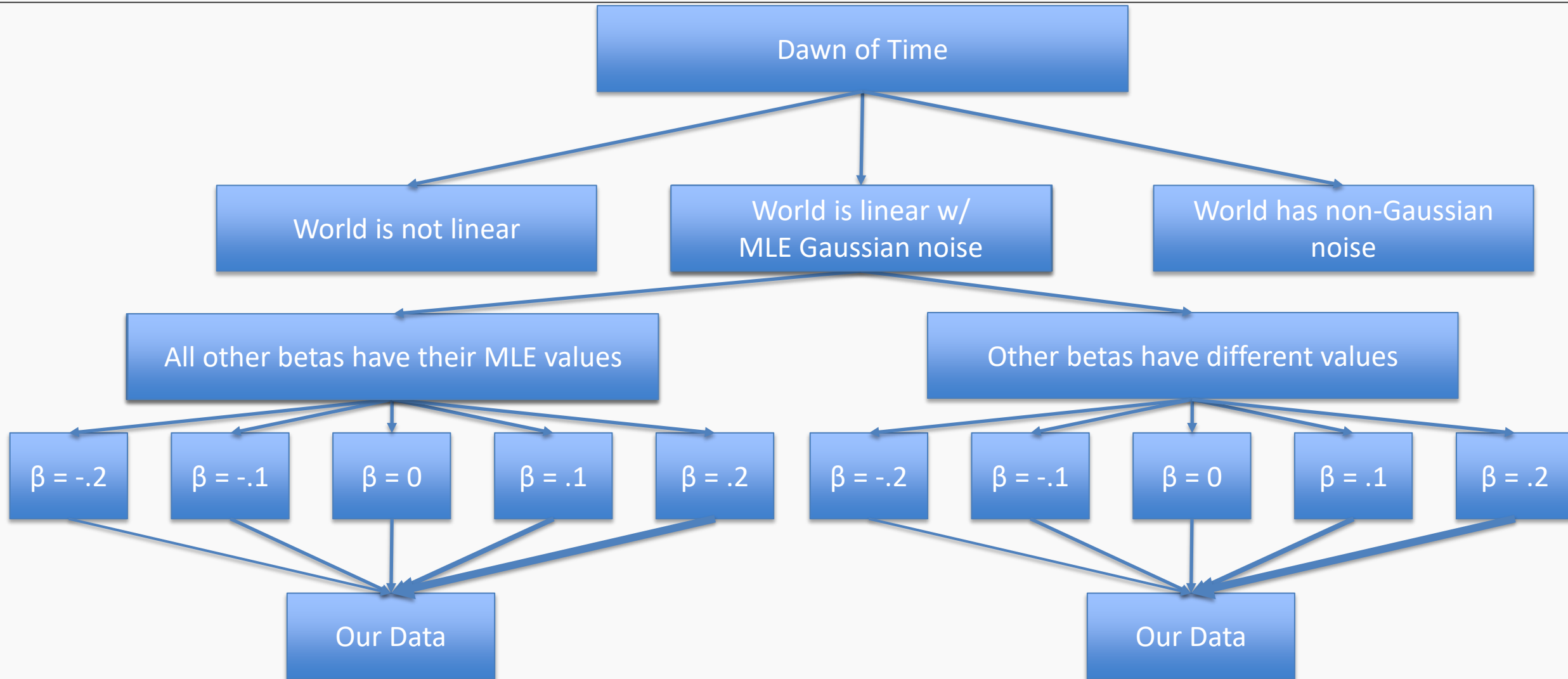


- What if we tested LOTS of possible values of beta?
- We end up with a set/interval of surviving values, e.g. [.1,.3]
- Since each beta was tested under the rule “reject this beta if the observed results are in the top 5% of weird datasets under this model”, we have [.1,.3] as a 95% confidence interval

HW Preview

- The 209 homework touches on another kind of confidence interval
 - Class: “How well have I estimated beta?”
 - HW: “How well can I estimate the *mean* response at each X ?”
 - Bonus: “How well can I estimate the *possible* responses at each X ”?

Remember those assumptions?



- We rejected the null model(s) as tested, not the *idea* that $\beta=0$ – assumptions matter

Review

- Ruling out a single model isn't much in general
- Sometimes, if we are lucky, ruling out a single model is enough to rule out a whole class of models
- Assumptions our model makes are weak points that should be justified and checked for accuracy

STATISTICS: REVIEW

You made it!

Review

- To test a particular model (a particular set of parameters) we must:
 1. Specify a data generating process
 2. Pick a way to measure whether our data plausibly comes from the process
 3. Pick a rule for when a model cannot be trusted (when is the range of simulated results too different from the observed data?)
- *What features make for a good test?*
 - We want to make as few assumptions as possible, and choose a measure that is sensitive to deviations from the model
 - If we're clever, we might get math that lets us skip simulating from the model
 - Tension: more assumptions make math easier, fewer assumptions make results broader
- There is no such thing as THE null hypothesis. It's only **A** null hypothesis.
 - A p value only tests one null hypothesis, and is rarely enough

Going forward

As the course moves on, we'll see

- Flexible assumptions about the data generating process
 - Generalized Linear Models
- Ways of making fewer assumptions about the data generating process:
 - Bootstrapping
 - Permutation tests
- Easier questions: Instead of 'find a model that explains the world', 'pick the model that predicts best'
 - Validation sets and cross validation

THANK YOU!