INFO 2950: Intro to Data Science

Lecture 12
2023-10-04

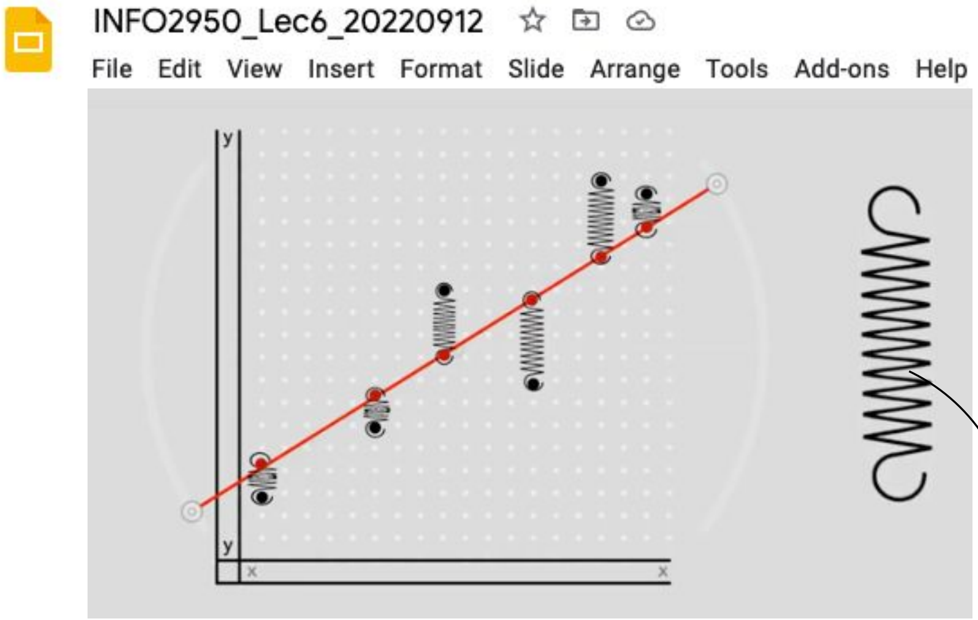
Agenda

1. Stochastic Gradient Descent
2. Interaction Effects
3. Rank Transformations
4. Admin

Attendance!

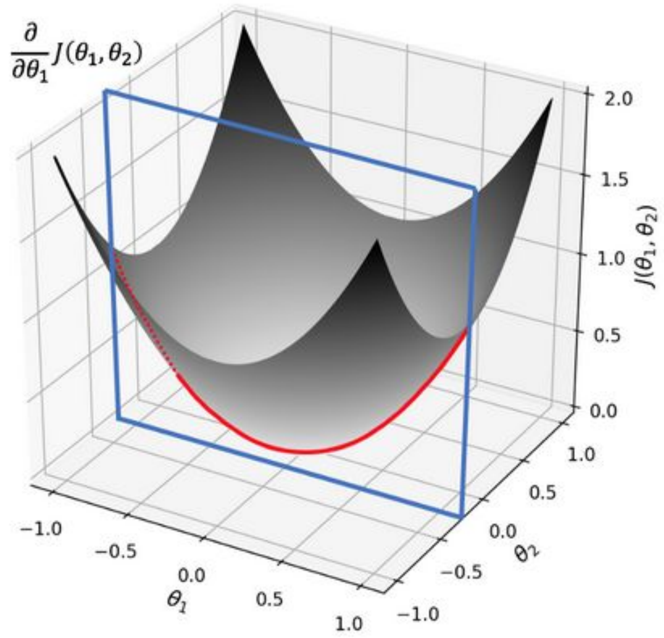
tinyurl.com/yzyr89y5

Least Squares Intuition for Single Variable Regression



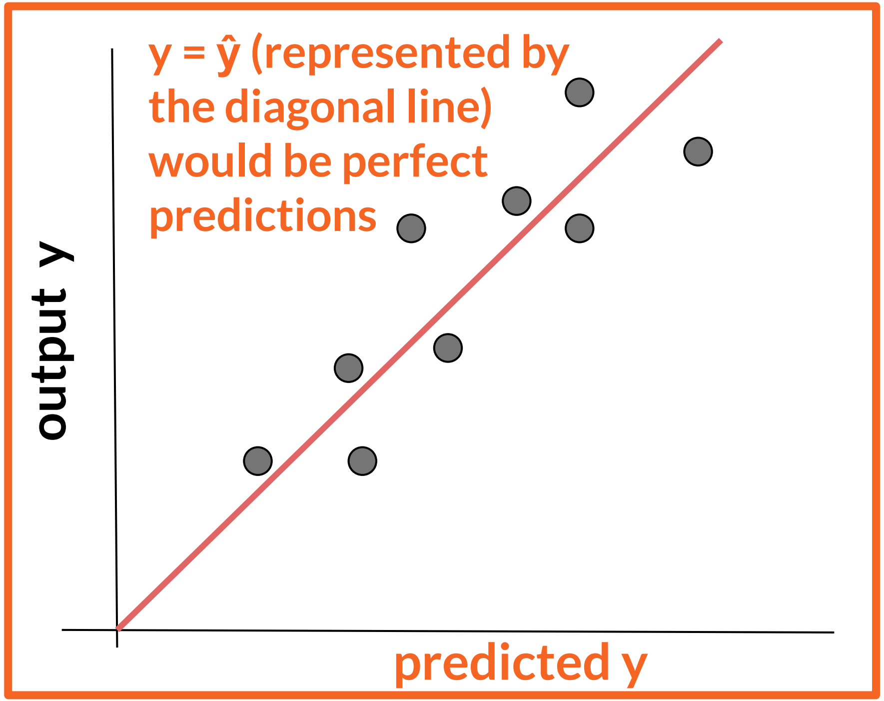
https://www.jmp.com/en_us/statistics-knowledge-portal/what-is-regression/the-method-of-least-squares.html

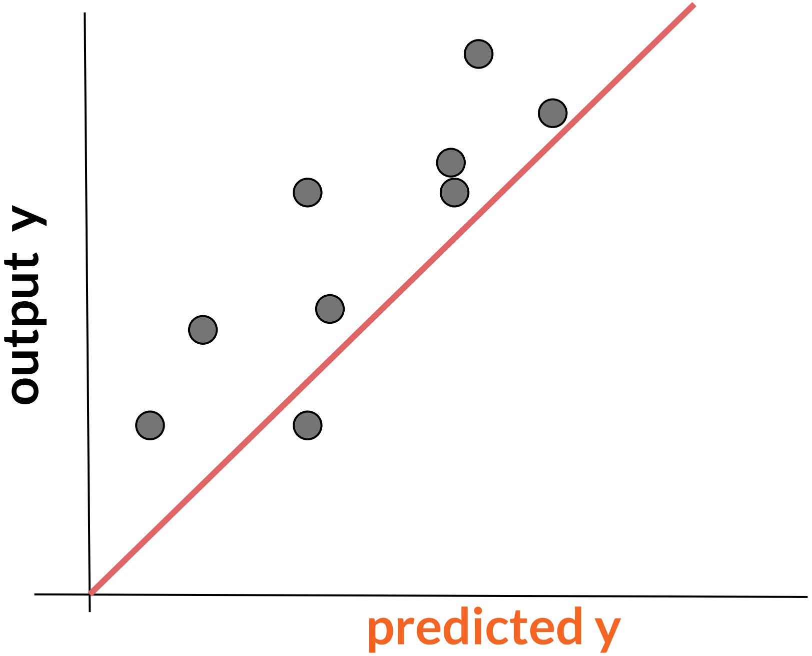
Multivariable Minimizing Intuition



https://towardsdatascience.com/machine-learning-bit-by-bit-multivariate-gradient-descent-e198fdd0df85

Which model has better predictions?





Experiment: be the optimizer

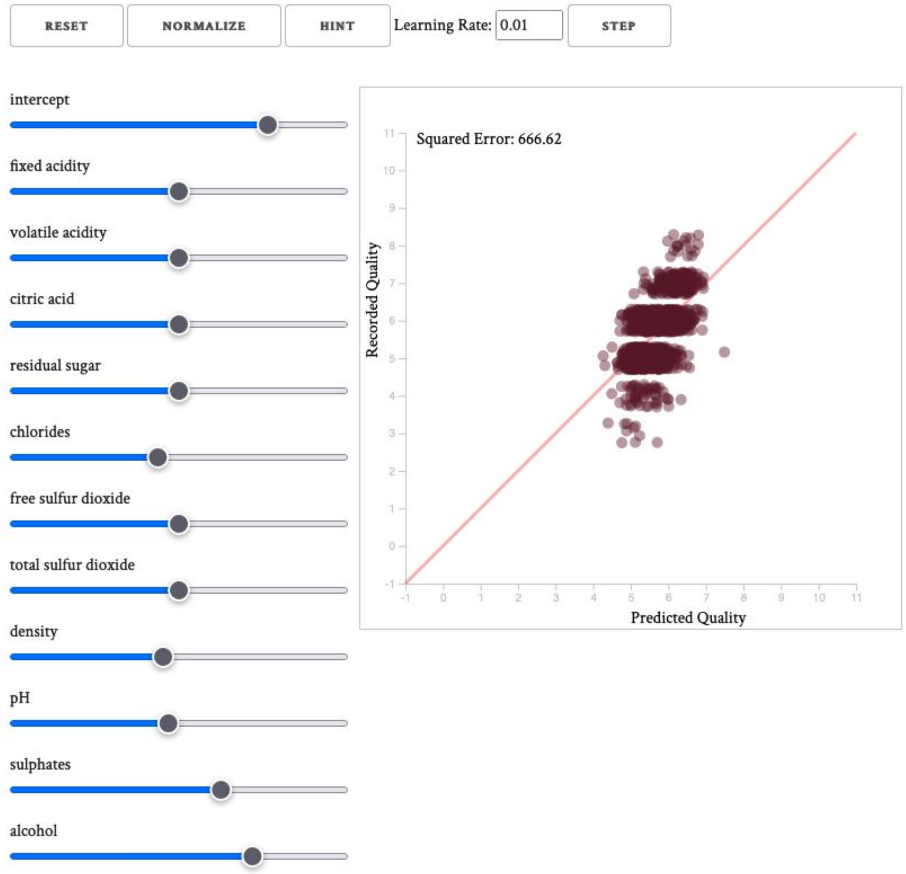
You will have a set of inputs (chemical properties of wines) and a set of outputs (rating of each wine)

For each input, you will need to set the β parameter to minimize the squared difference between your **predicted rating and the actual rating**

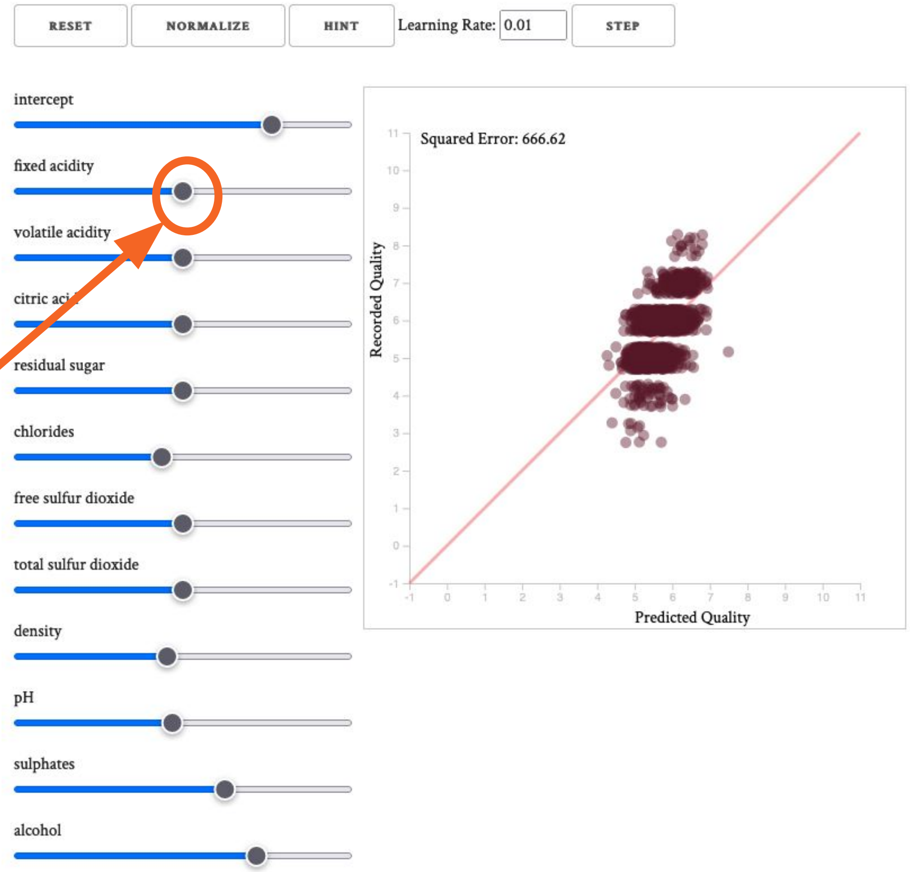
Move the slider for each variable left or right to change parameter settings

Each of these is a coefficient in a regression

Quality ~ intercept + fixed_acidity + volatile_acidity + …



How do we know this is the right coefficient for fixed acidity, instead of somewhere to the left or right of it?



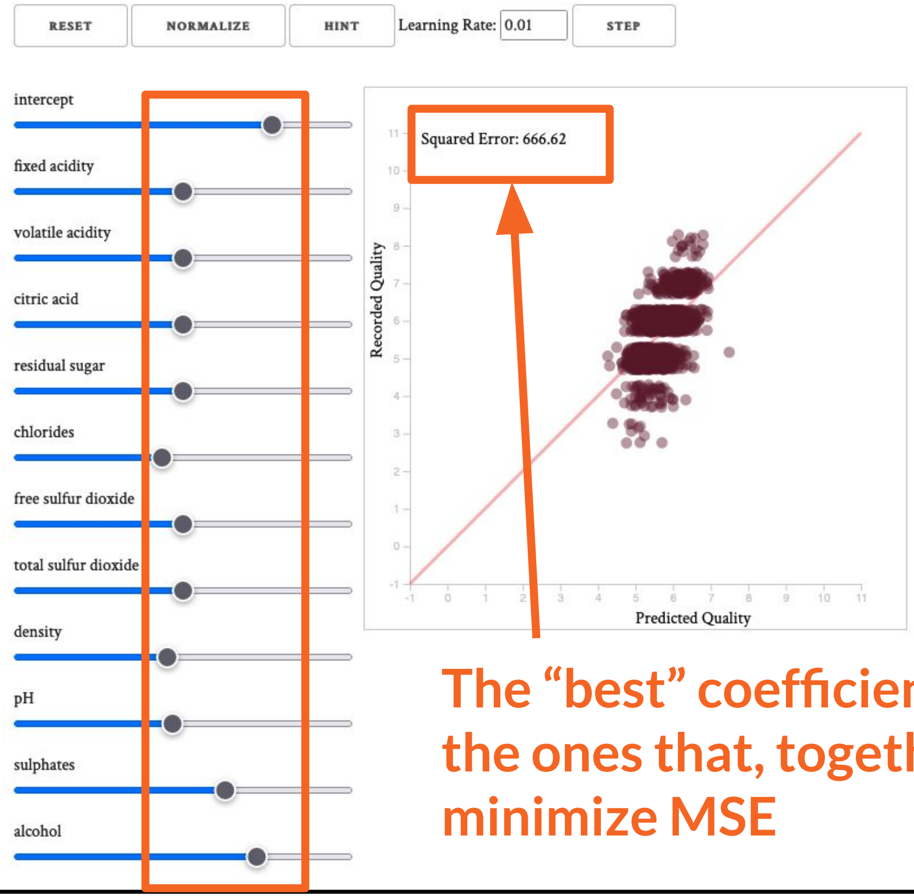
https://mimno.infosci.cornell.edu/info2950/interactive/wine.html

Multivariable Gradient Intuition

https://tinyurl.com/mr3hphey

1. What is the smallest squared error you can find?

2. Which sliders move the points more or less?



The "best" coefficients are the ones that, together, minimize MSE

https://mimno.infosci.cornell.edu/info2950/interactive/wine.html

Why is this hard?

```
"fixed acidity";"volatile acidity";"citric
acid";"residual sugar";"chlorides";"free sulfur
dioxide";"total sulfur
dioxide";"density";"pH";"sulphates";"alcohol";"quality"
7.4;0.7;0;1.9;0.076;11;34;0.9978;3.51;0.56;9.4;5
7.8;0.88;0;2.6;0.098;25;67;0.9968;3.2;0.68;9.8;5
7.8;0.76;0.04;2.3;0.092;15;54;0.997;3.26;0.65;9.8;5
11.2;0.28;0.56;1.9;0.075;17;60;0.998;3.16;0.58;9.8;6
```

Why is this hard?

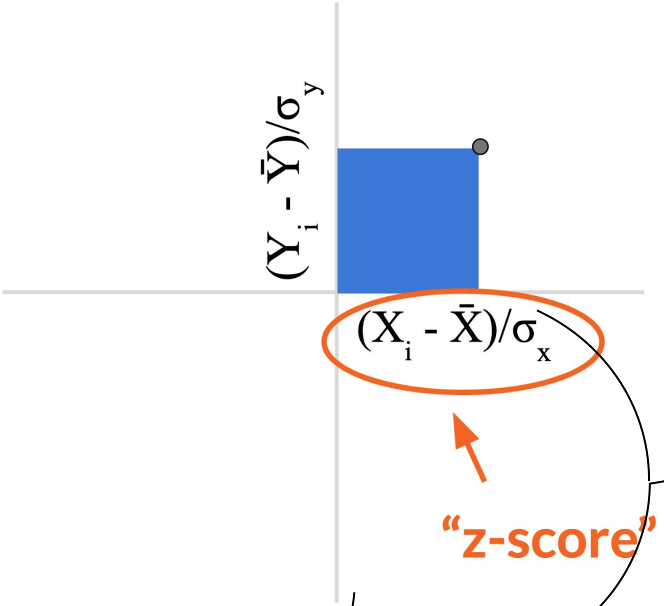
```
"fixed acidity";"volatile acidity";"citric
acid";"residual sugar";"chlorides";"free sulfur
dioxide";"total sulfur
dioxide";"density";"pH";"sulphates";"alcohol";"quality"
7.4;0.7;0;1.9;0.076;11;34;0.9978;3.51;0.56;9.4;5
7.8;0.88;0;2.6;0.098;25;67;0.9968;3.2;0.68;9.8;5
7.8;0.76;0.04;2.3;0.092;15;54;0.997;3.26;0.65;9.8;5
11.2;0.28;0.56;1.9;0.075;17;60;0.998;3.16;0.58;9.8;6
```

Why is this hard?

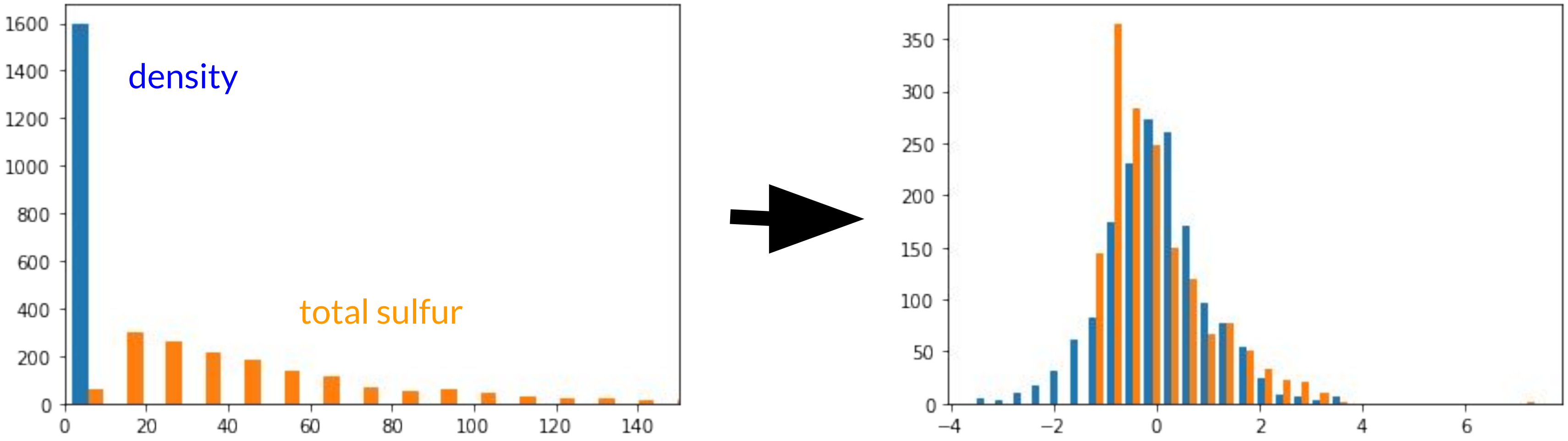
```
"fixed acidity";"volatile acidity";"citric
acid";"residual sugar";"chlorides";"free sulfur
dioxide";"total sulfur
dioxide";"density";"pH";"sulphates";"alcohol";"quality"
7.4;0.7;0;1.9;0.076;11;34;0.9978;3.51;0.56;9.4;5
7.8;0.88;0;2.6;0.098;25;67;0.9968;3.2;0.68;9.8;5
7.8;0.76;0.04;2.3;0.092;15;54;0.997;3.26;0.65;9.8;5
11.2;0.28;0.56;1.9;0.075;17;60;0.998;3.16;0.58;9.8;6
```

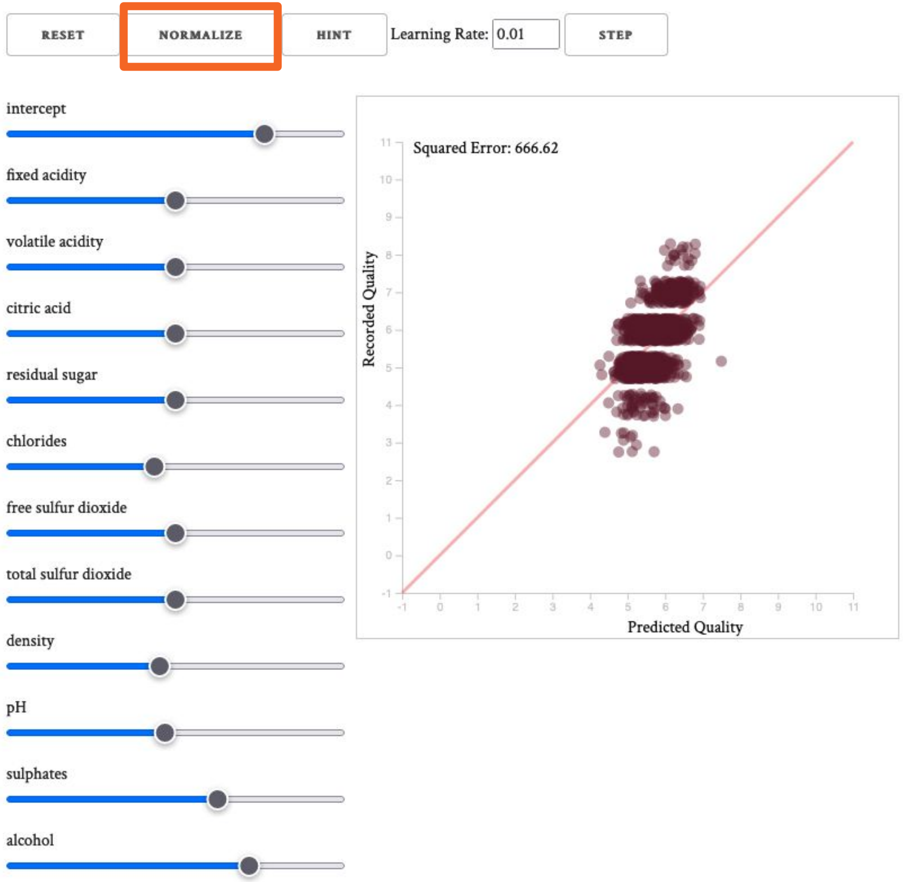
Subtracting the mean and dividing by the std. dev. *normalizes* the variables

Values are now comparable, regardless of units



Input variables are on different scales: *normalize* with z-scores



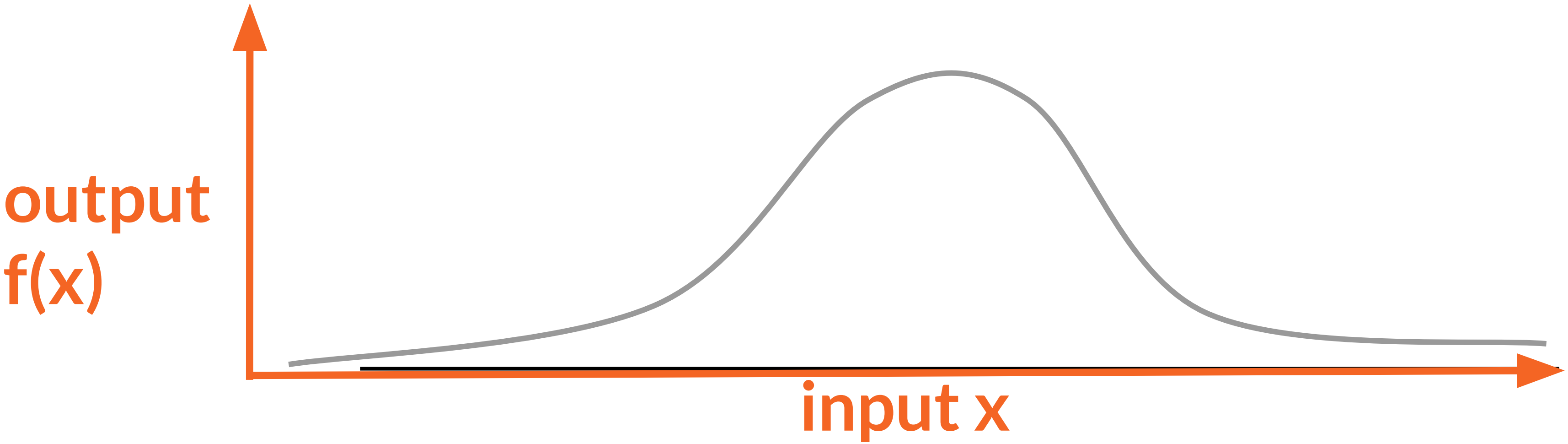


https://mimno.infosci.cornell.edu/info2950/interactive/wine.html

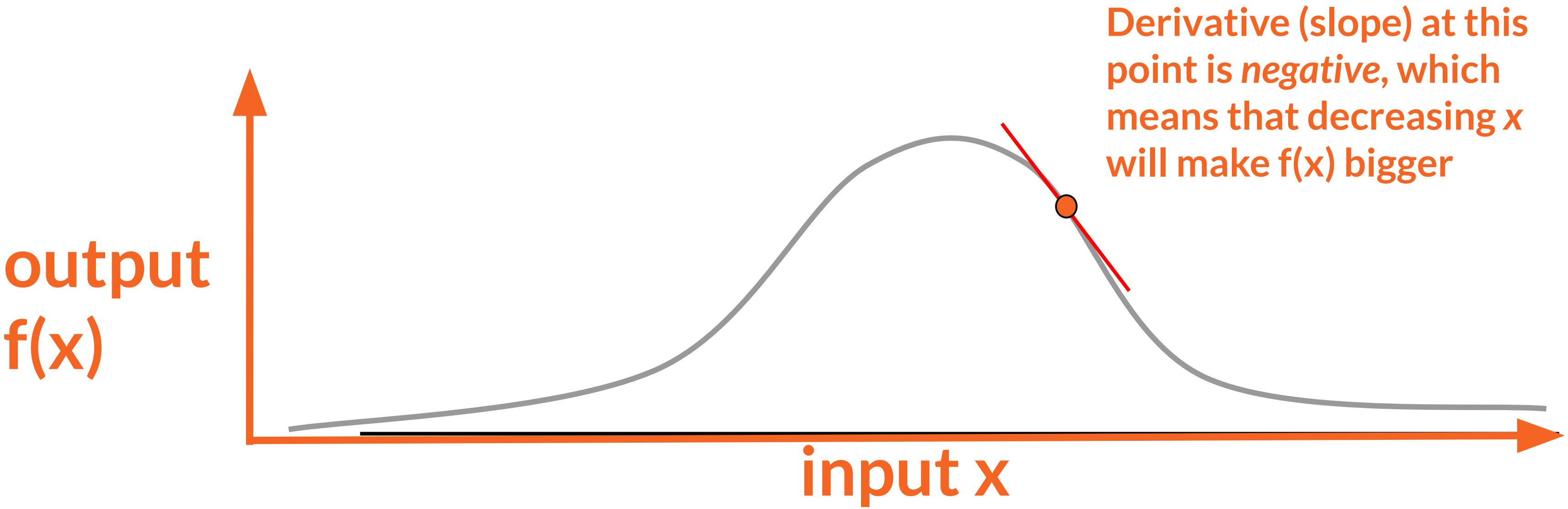
How do you minimize squared error?

Before, we talked about using calculus to do this by hand (for single variable linear regressions)

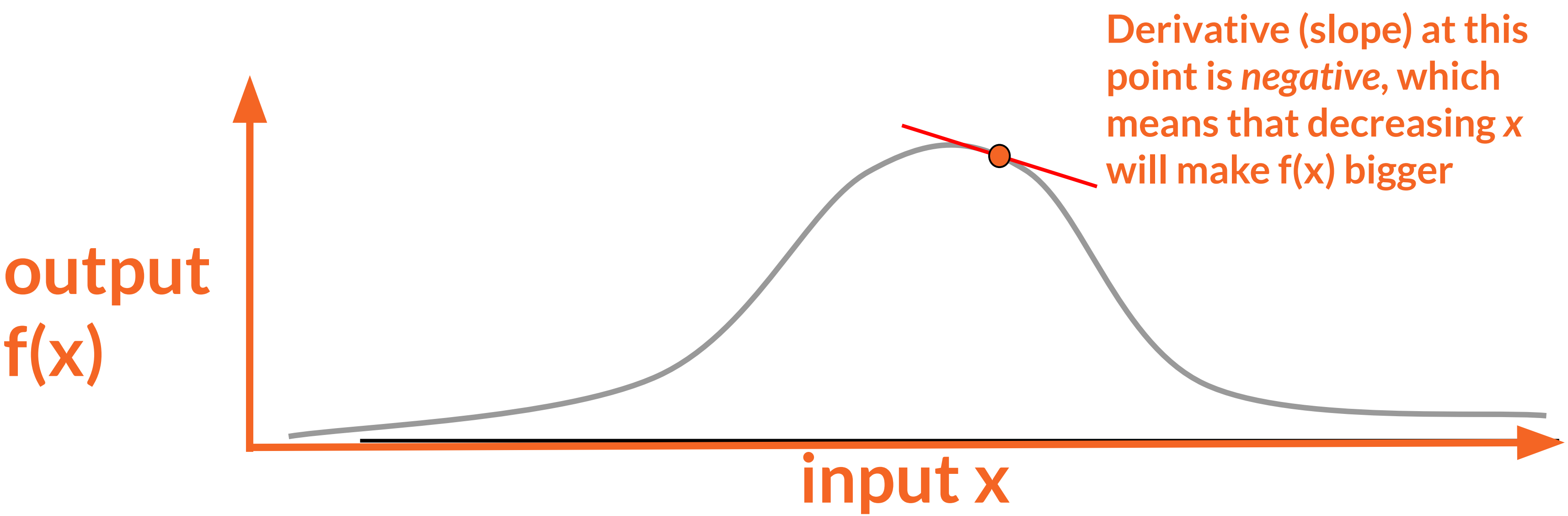
What is the maximum value of output f(x)?



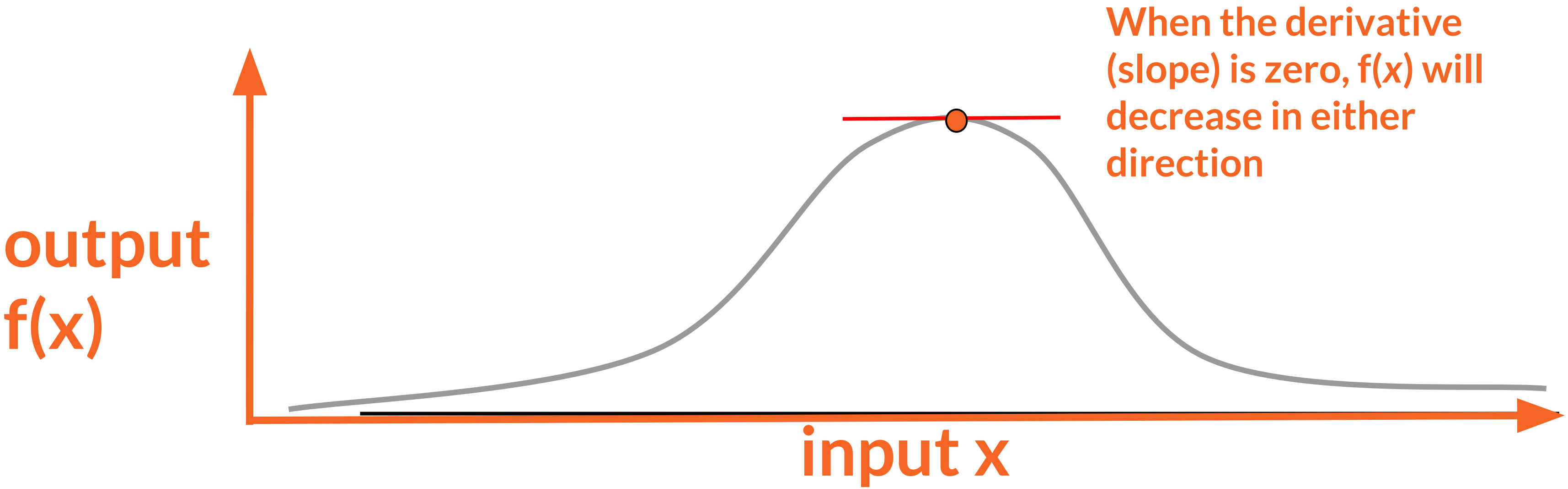
The derivative is a hint: which way would you go to increase f(x)?



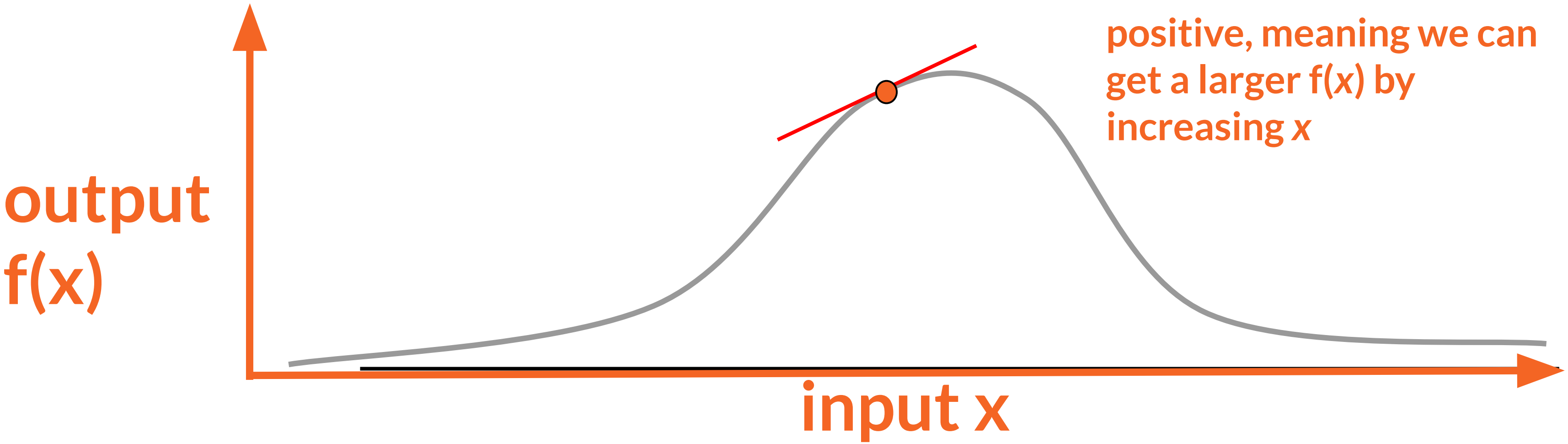
The derivative is a hint: which way would you go to increase f(x)?



The derivative is a hint: which way would you go to increase f(x)?



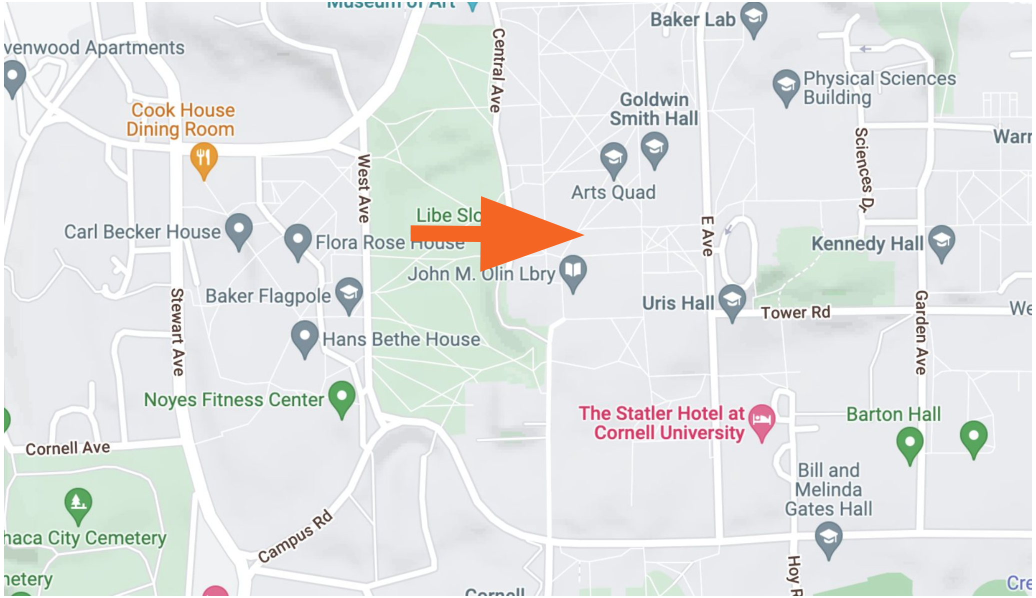
The derivative is a hint: which way would you go to increase f(x)?



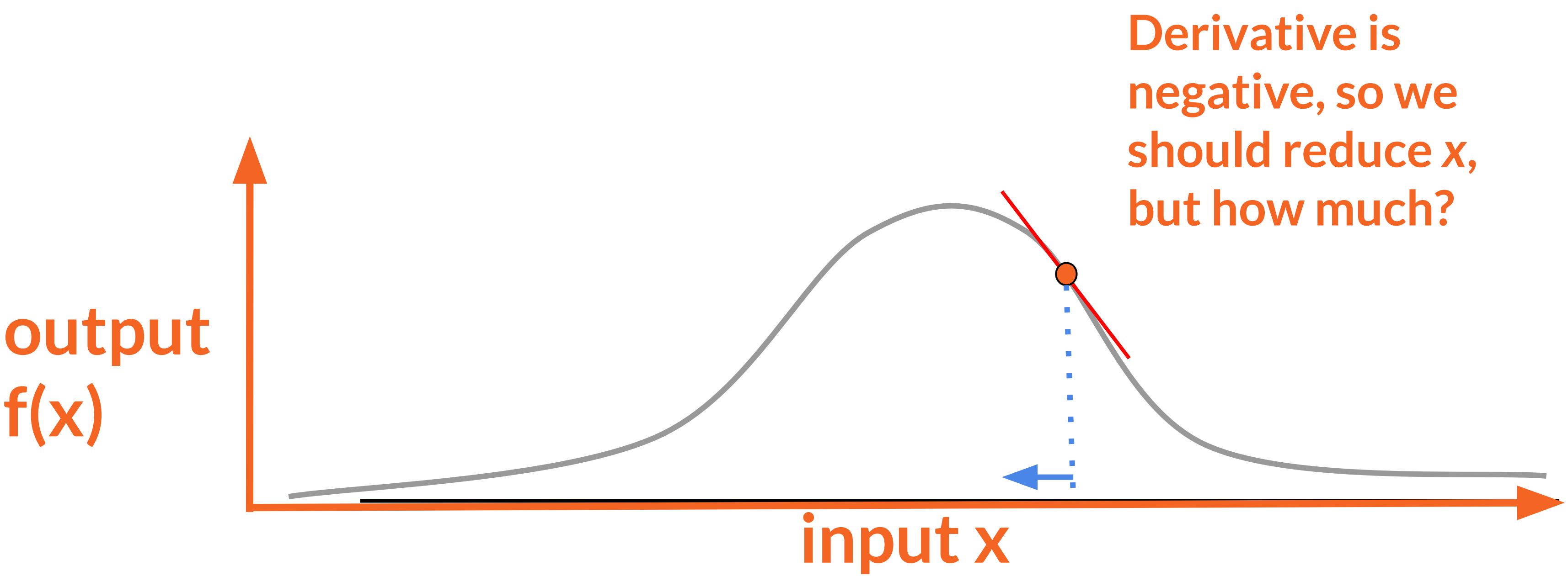
If we go too far, the derivative will be positive, meaning we can get a larger f(*x*) by increasing *x*

Gradient is a hint in multiple directions

The slope is steep in the East-West direction, but flat in the North-South direction



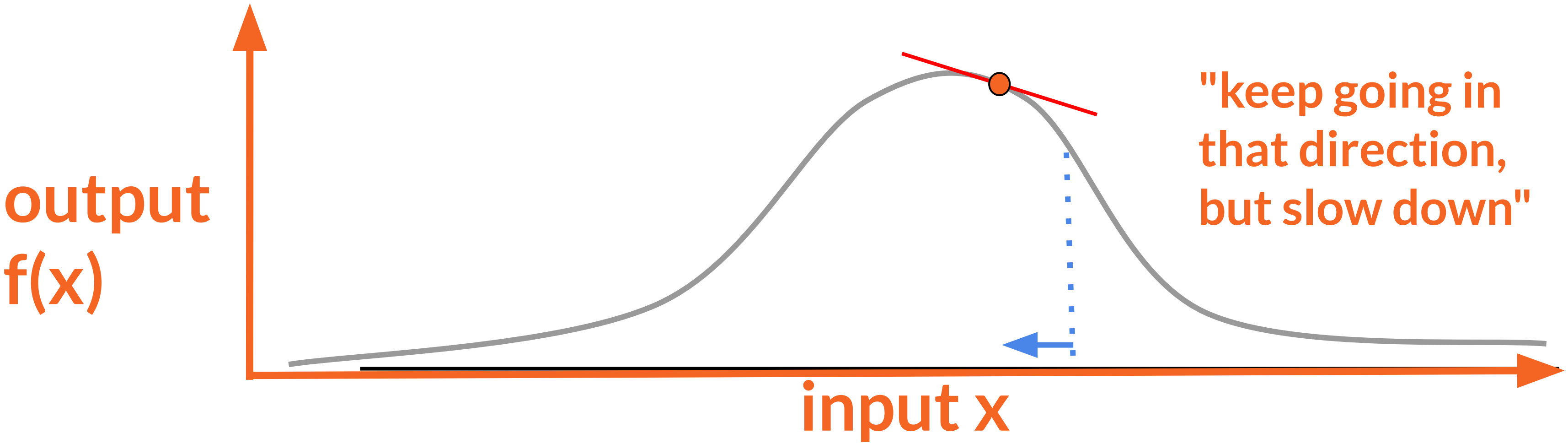
Stochastic Gradient Descent (SGD)



SGD: move β in a direction specified by the gradient

Slope still negative, but less so:

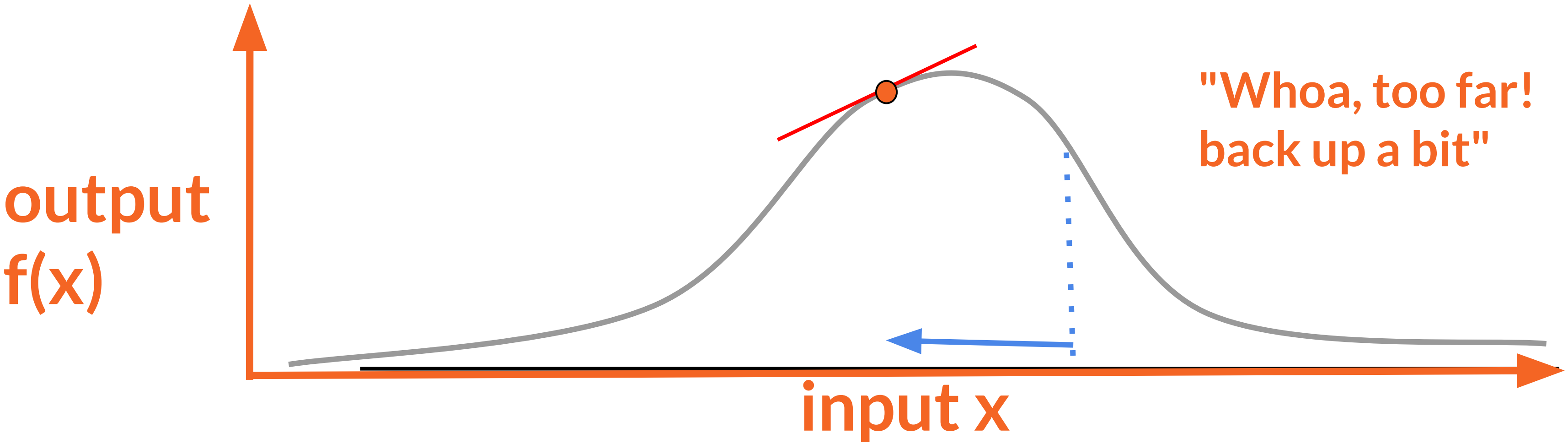
"keep going in that direction, but slow down"



SGD: "step size" or "learning rate" is important

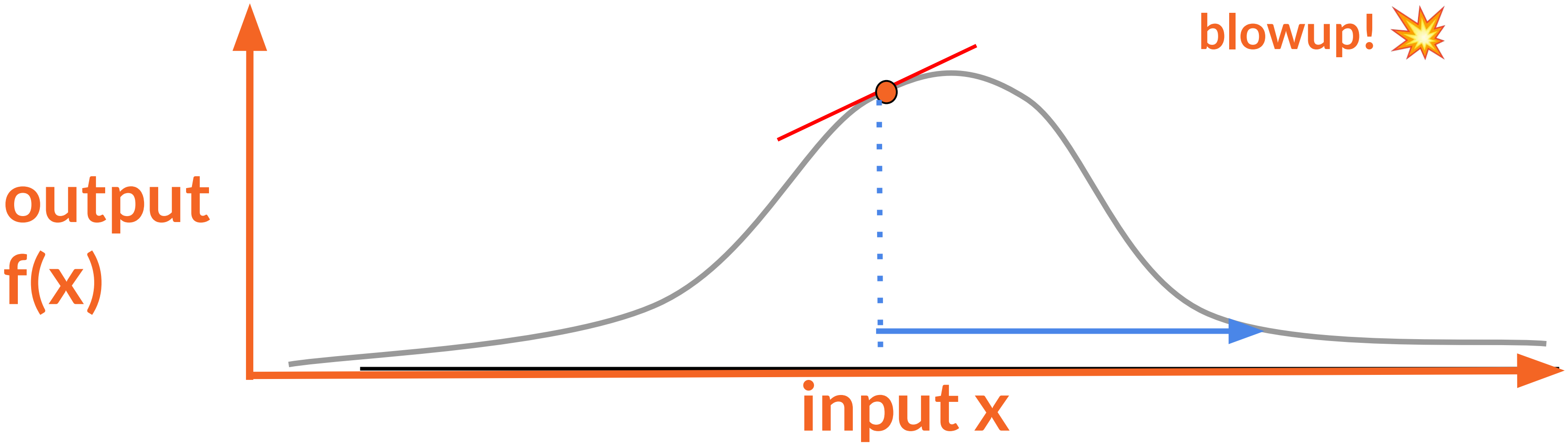
If you take a big step you may overshoot:

"Whoa, too far! back up a bit"



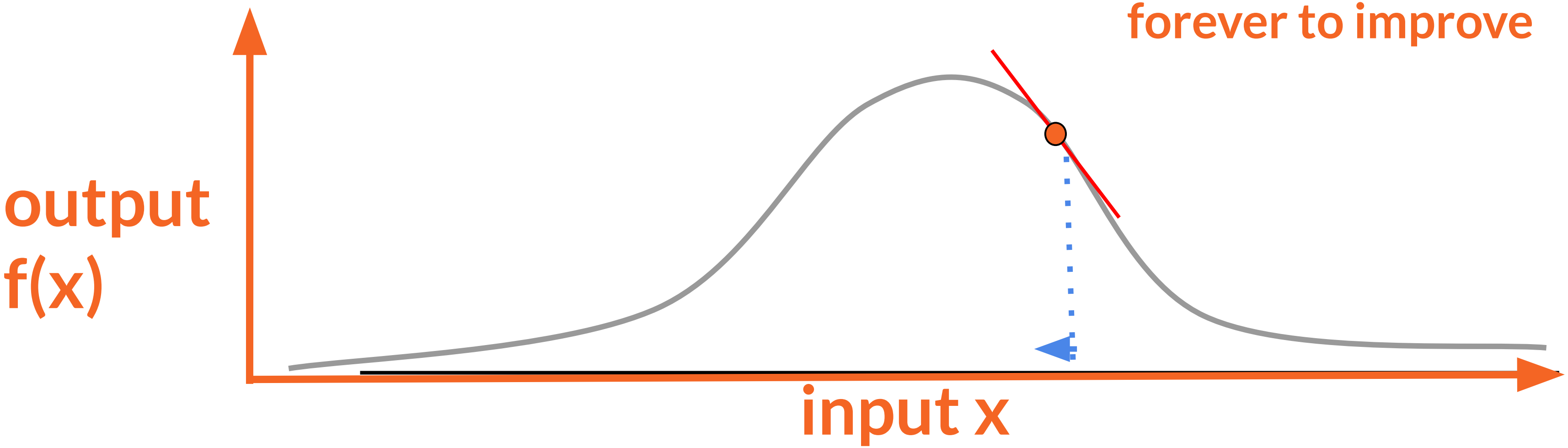
SGD: "step size" or "learning rate" is important

Overshooting can lead to oscillation and blowup! 💥

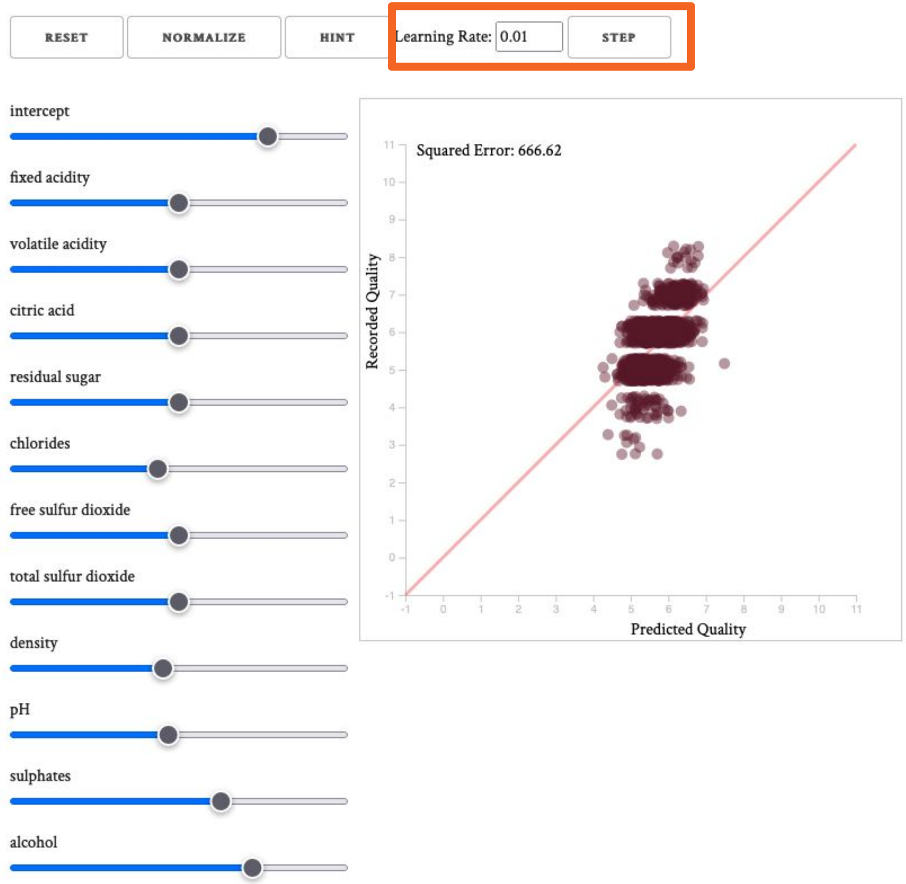


SGD: "step size" or "learning rate" is important

Too small a step and it will take forever to improve

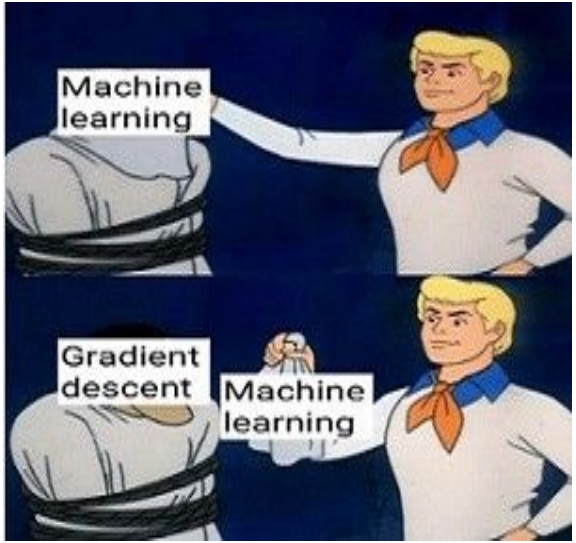


https://www.memecreator.org/meme/global-optimum-sgd/



https://mimno.infosci.cornell.edu/info2950/interactive/wine.html

Takeaways on gradients



- "Stochastic gradient descent" is used to find minima / maxima for complicated models (e.g. multivariable regression)
  - choose a **learning** rate to do this efficiently
  - this is the core of modern machine **learning**!

https://towardsdatascience.com/batch-mini-batch-stochastic-gradient-descent-7a62ecba642a

Admin: Extra Credit! 🌟

- Two surveys, +10 points towards HW2 grade for filling out each:
  - Mid-Semester Course Feedback
  - Midterm TA Evaluations
- Surveys due on Oct 13

Regressions

- We’ve covered how to “**fit**” given a df that has x’s and y:
  - LinearRegression
  - LogisticRegression
- After fitting the models, we’ve covered how to **interpret** the models based on coefficients

Regressions

- We’ve covered how to “**fit**” given a df that has x’s and y:
  - LinearRegression
  - LogisticRegression
- After fitting the models, we’ve covered how to **interpret** the models based on coefficients
  - Mostly!

Variables can interact

"Vitamin C and vitamin E are possibly the best example of a skincare power couple as they each enhance the effects of the other. For example, vitamin C helps to regenerate vitamin E, making it more readily available to protect the skin from free radicals. **Vitamin E returns the favor by increasing the action of vitamin C four-fold (4x)** [19][20]."

https://www.sciencebecomesher.com/skincare-ingredients-that-work-better-together/

Variables can interact

"Vitamin C and vitamin E are possibly the best example of a skincare power couple as they each enhance the effects of the other. For example, vitamin C helps to regenerate vitamin E, making it more readily available to protect the skin from free radicals. **Vitamin E returns the favor by increasing the action of vitamin C four-fold (4x)** [19][20]."

Results are *not* linear in C + E!

https://www.sciencebecomesher.com/skincare-ingredients-that-work-better-together/

Multivariable Regressions

- Our interpretations (both for linear regression and logistic regression) have assumed that:
  - **increasing x_1 by one unit only affects y, so the only source of an effect due to x_1 is captured by β_1**

Multivariable Regressions

- Our interpretations (both for linear regression and logistic regression) have assumed that:
  - **increasing x_1 by one unit only affects y, so the only source of an effect due to x_1 is captured by β_1**

- What if this isn't the case? What if changing $\mathbf{x_1}$ also has something to do with $\mathbf{x_2}$?

Introducing: Interactions

- "Interactions" in regression math are when different x's should be considered *together*

- **Interactions** (in math) ≈ **Intersectionality** (in life)

DeGraffenreid v. General Motors (1976)

- GM did not hire Black women before 1964
- In the early 1970s recession, GM did layoffs by seniority → all the Black women were laid off
- 5 Black women sued GM over discrimination by gender *and* race
- Unsuccessful because the court didn't know how to deal with the *intersection* of gender and race

https://www.vox.com/the-highlight/2019/5/20/18542843/intersectionality-conservatism-law-race-gender-discrimination

DeGraffenreid v. General Motors (1976)

> "The legislative history surrounding Title VII does not indicate that the goal of the statute was to create a new classification of 'black women' who would have greater standing than, for example, a black male. The prospect of the creation of new classes of protected minorities, governed only by the mathematical principles of permutation and combination, clearly raises the prospect of opening the hackneyed Pandora's box."

- Judge Harris Wangelin's ruling against the plaintiffs

https://www.vox.com/the-highlight/2019/5/20/18542843/intersectionality-conservatism-law-race-gender-discrimination

Multivar Regression: Interactions

- **Interactions** (in math) ≈ **Intersectionality** (in life)
- y = teaching evaluations

Keep an eye out for mid-semester teaching evaluations for TAs!

They've worked really hard and any praise for them via feedback would be appreciated :)

Multivar Regression: Interactions

- **Interactions** (in math) ≈ **Intersectionality** (in life)

- y = teaching evaluations
- x_1 = instructor race
- x_2 = instructor gender

https://link.springer.com/article/10.1057/s41302-020-00176-2

Multivar Regression: Interactions

- **Interactions** (in math) ≈ **Intersectionality** (in life)

- y = teaching evaluations
- x_1 = instructor race
- x_2 = instructor gender

Use binary variables here to teach this concept

https://link.springer.com/article/10.1057/s41302-020-00176-2

Multivar Regression: Interactions

- **Interactions** (in math) ≈ **Intersectionality** (in life)

- y = teaching evaluations
- x_1 = instructor race != white
- x_2 = instructor gender != male

https://link.springer.com/article/10.1057/s41302-020-00176-2

Multivar Regression: Interactions

- y = teaching evaluations
- x_1 = instructor race
- x_2 = instructor gender

- Evaluations are worse for non-white instructors
- Evaluations are worse for female instructors
- Evaluations can be ***disproportionately worse*** for female non-white instructors

https://link.springer.com/article/10.1057/s41302-020-00176-2

Multivar Regression: Interactions

- y = teaching evaluations
- x_1 = instructor race != white
- x_2 = instructor gender != male

- $y \sim x_1 + x_2 + x_1 * x_2$

the product of two covariates (which can include dummies) forms an "interaction term"

Multivar Regression: Interactions

- y = teaching evaluations
- x_1 = instructor race != white
- x_2 = instructor gender != male

- $y \sim x_1 + x_2 + x_1{}^*x_2$

4 possibilities this can take:
1. Non-white non-male
2. White non-male
3. Non-white male
4. White male

Multivar Regression: Interactions

- y = teaching evaluations
- x_1 = instructor race != white
- x_2 = instructor gender != male

- Non-white non-male → $x_1 = 1, x_2 = 1$
- Non-white male → $x_1 = 1, x_2 = 0$
- White non-male → $x_1 = 0, x_2 = 1$
- White male → $x_1 = 0, x_2 = 0$

Multivar Regression: Interactions

- y = teaching evaluations
- x_1 = instructor race != white
- x_2 = instructor gender != male

- $y \sim x_1 + x_2 + x_1{}^*x_2$

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_{1*2} x_1{}^* x_2$$

Extra coefficient can be estimated for the combination

Multivar Regression: Interactions

- y = teaching evaluations
- x_1 = instructor race != white
- x_2 = instructor gender != male

- $y \sim x_1 + x_2 + x_1 * x_2$ **(made up numbers below)**
- $y = 4.5 - 0.4x_1 - 0.5x_2 - 0.1x_1 * x_2$

Fill in the table

- $y = 4.5 - 0.4x_1 - 0.5x_2 - 0.1x_1 * x_2$
- Fill out last column in terms of **α, β_1, β_2, β_{1*2}**

Demographic	x_1	x_2	Example	Value	Equation
White male	0	0	ŷ = 4.5 - 0.4*0 - 0.5*0 - 0.1*0*0	4.5	ŷ = ________
Non-white male	1	0	ŷ = 4.5 - 0.4*1 - 0.5*0 - 0.1*1*0	4.1	ŷ = ________
White non-male	0	1	ŷ = 4.5 - 0.4*0 - 0.5*1 - 0.1*0*1	4.0	ŷ = ________
Non-white non-male	1	1	ŷ = 4.5 - 0.4*1 - 0.5*1 - 0.1*1*1	3.5	ŷ = ________

Multivar Regression: Interactions

- $y = 4.5 - 0.4x_1 - 0.5x_2 - 0.1x_1{*}x_2$
- For categorical interactions, coefficients are **additive**

Demographic	x_1	x_2	Example	Value	Equation
White male	0	0	$\hat{y} = 4.5 - 0.4*0 - 0.5*0 - 0.1*0*0$	4.5	$\hat{y} = \alpha$
Non-white male	1	0	$\hat{y} = 4.5 - 0.4*1 - 0.5*0 - 0.1*1*0$	4.1	$\hat{y} = \alpha + \beta_1$
White non-male	0	1	$\hat{y} = 4.5 - 0.4*0 - 0.5*1 - 0.1*0*1$	4.0	$\hat{y} = \alpha + \beta_2$
Non-white non-male	1	1	$\hat{y} = 4.5 - 0.4*1 - 0.5*1 - 0.1*1*1$	3.5	$\hat{y} = \alpha + \beta_1 + \beta_2 + \beta_{1*2}$

Multivar Regression: Interactions

- $y = 4.5 - 0.4x_1 - 0.5x_2 - 0.1x_1{}^*x_2$
- **If you're at the intersection (x_1=x_2=1), you have an *extra* β_{1*2} added to your ŷ**

Demographic	**x_1**	**x_2**	**Example**	**Value**	**Equation**
White male	**0**	**0**	$\hat{y} = 4.5 - 0.4*\mathbf{0} - 0.5*\mathbf{0} - 0.1*\mathbf{0}*\mathbf{0}$	4.5	$\hat{y} = \alpha$
Non-white male	**1**	**0**	$\hat{y} = 4.5 - 0.4*\mathbf{1} - 0.5*\mathbf{0} - 0.1*\mathbf{1}*\mathbf{0}$	4.1	$\hat{y} = \alpha + \beta_1$
White non-male	**0**	**1**	$\hat{y} = 4.5 - 0.4*\mathbf{0} - 0.5*\mathbf{1} - 0.1*\mathbf{0}*\mathbf{1}$	4.0	$\hat{y} = \alpha + \beta_2$
Non-white non-male	**1**	**1**	$\hat{y} = 4.5 - 0.4*\mathbf{1} - 0.5*\mathbf{1} - 0.1*\mathbf{1}*\mathbf{1}$	3.5	$\hat{y} = \alpha + \beta_1 + \beta_2 + \beta_{1*2}$

Interpreting Regressions

- If **no interactions**, interpret each different x_i separately
  - Summarize Relationship | Predict Outcome | Outliers & Oddities
- If **have interactions**, interpret by plugging in values for different combinations of x_i
  - Predict Outcome | Outliers & Oddities

Interpreting Regressions

- If **no interactions**, interpret each different x_i separately
  - Summarize Relationship | Predict Outcome | Outliers & Oddities
- If **have interactions**, interpret by plugging in values for different combinations of x_i
  - Predict Outcome | Outliers & Oddities

Multivar Regression: Interactions

- Our model predicts that non-white non-male instructors are rated lowest by student evaluations at 3.5, which is a full point lower than white male instructors, who we predict to have the highest student evaluations at 4.5.

Demographic	x_1	x_2	Example	Value	Equation
White male	0	0	$y^{hat} = 4.5 - 0.4*0 - 0.5*0 - 0.1*0*0$	4.5	$y^{hat} = \alpha$
Non-white male	1	0	$y^{hat} = 4.5 - 0.4*1 - 0.5*0 - 0.1*1*0$	4.1	$y^{hat} = \alpha + \beta_1$
White non-male	0	1	$y^{hat} = 4.5 - 0.4*0 - 0.5*1 - 0.1*0*1$	4.0	$y^{hat} = \alpha + \beta_2$
Non-white non-male	1	1	$y^{hat} = 4.5 - 0.4*1 - 0.5*1 - 0.1*1*1$	3.5	$y^{hat} = \alpha + \beta_1 + \beta_2 + \beta_{1*2}$

Recall that interactions are like mathematical intersectionality.

Does it make sense to use interactions on the season dummies summer (x_2), fall (x_3), and/or winter (x_4)?

Dummies x Interactions

Does it make sense to use interactions on the season dummies x_2, x_3, and/or x_4?

No, because the seasons are mutually exclusive: $x_2 * x_3 = 0$, $x_2 * x_4 = 0$, $x_3 * x_4 = 0$. Including a 0 term in your regression does not add any meaning to your model!

Can’t interact variables that never have interaction!

y	x_1			x_2	x_3	x_4
Temp (F)	**Pressure**	**Season**	**Spring**	**Summer**	**Fall**	**Winter**
80	81	Summer	0	1	0	0
50	63	Fall	0	0	1	0
70	75	Spring	1	0	0	0
…	…	…	…	…	…	…

Interpreting interactions

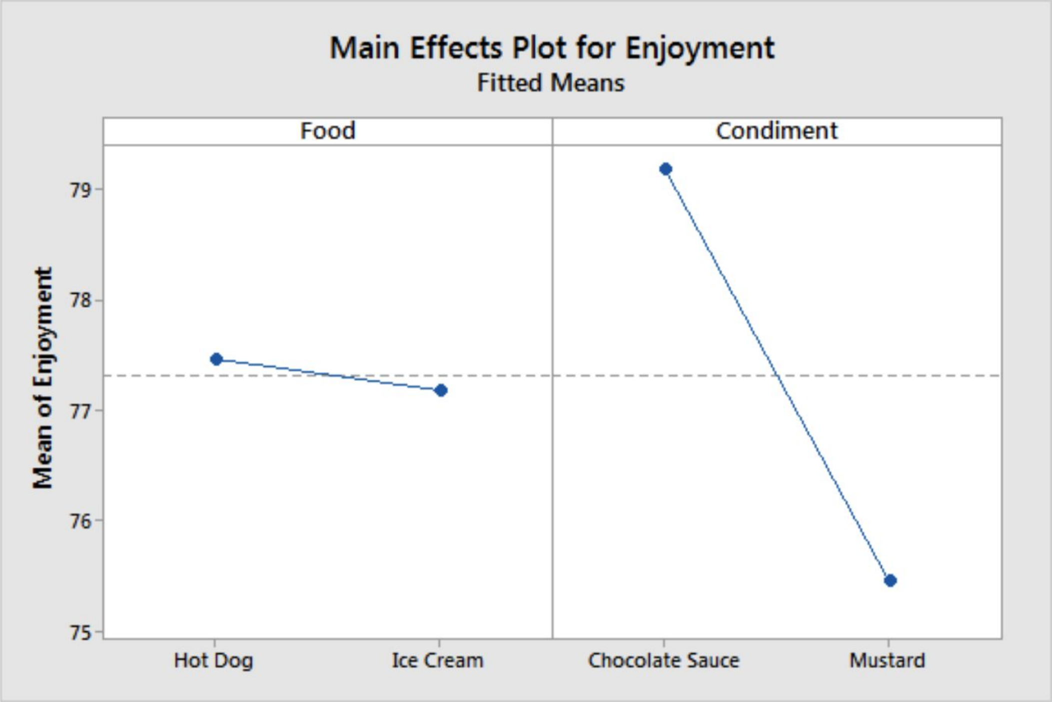
- y = enjoyment of food given
  - two types of food [hot dog; ice cream] and
  - two types of condiments [mustard; chocolate sauce]

- x_1 = **food == hot dog**

- x_2 = **condiment == chocolate sauce**

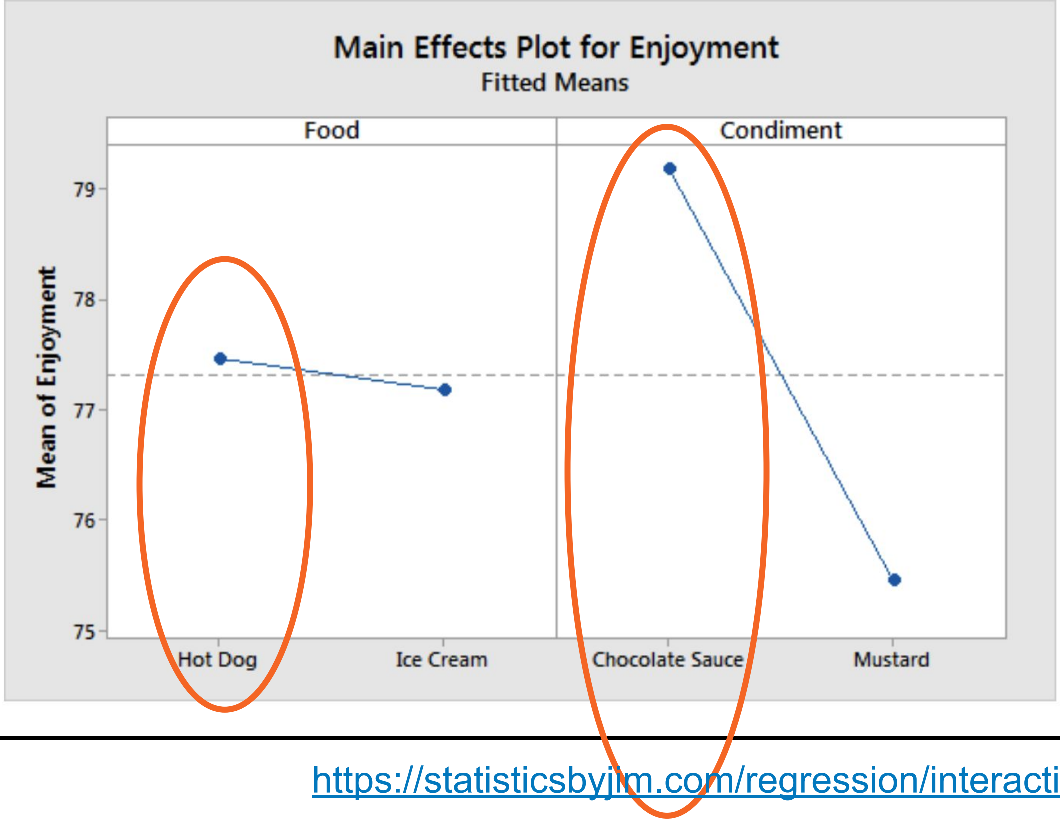
https://statisticsbyjim.com/regression/interaction-effects/

If you pick items with highest *y* means…



https://statisticsbyjim.com/regression/interaction-effects/

If you pick items with highest *y* means...



https://statisticsbyjim.com/regression/interaction-effects/

Interpreting Regressions

- If **no interactions**, interpret each different x_i separately
    - Summarize Relationship | Predict Outcome | Outliers & Oddities
- If **have interactions**, **interpret by plugging in values for different combinations of x_i**
    - Predict Outcome | Outliers & Oddities

Interpret Interactions

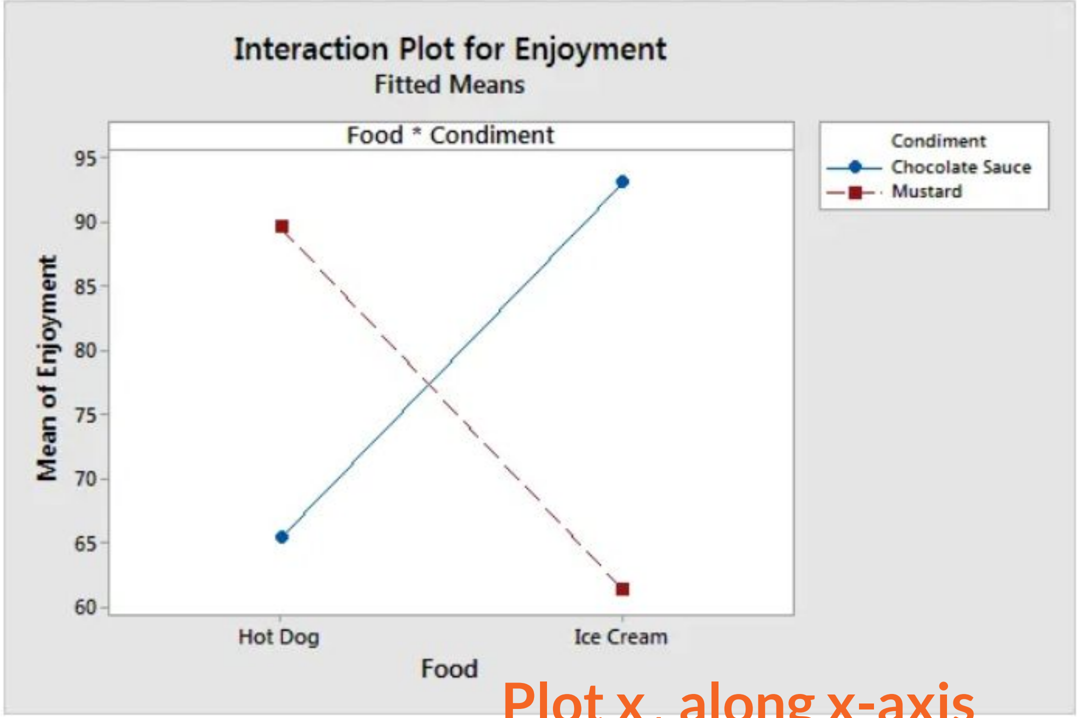
- y = enjoyment of food (on 100-point scale) among [hot dog; ice cream] and [mustard; chocolate sauce]
- x_1 = food == hot dog
- x_2 = condiment == chocolate sauce
- **$y = 61.3 + 28.3x_1 + 31.7x_2 - 56.0x_1 * x_2$**
- **How do you interpret this?**

Interpret Interactions

- y = enjoyment of food among [hot dog; ice cream] and [mustard; chocolate sauce]
- x_1 = food == hot dog
- x_2 = condiment == chocolate sauce

- $\mathbf{y = 61.3 + 28.3x_1 + 31.7x_2 - 56.0x_1{*}x_2}$
- **We expect ice cream with mustard to yield the lowest (61.3) enjoyment, ice cream with chocolate sauce to yield 93 enjoyment, hot dog with mustard to yield 89.6 enjoyment, and hot dog with chocolate sauce to yield 65.3 enjoyment. We propose bringing ice cream with chocolate sauce to the picnic to maximize enjoyment.**

Visual aid: interaction plot



Plot x_1 along x-axis

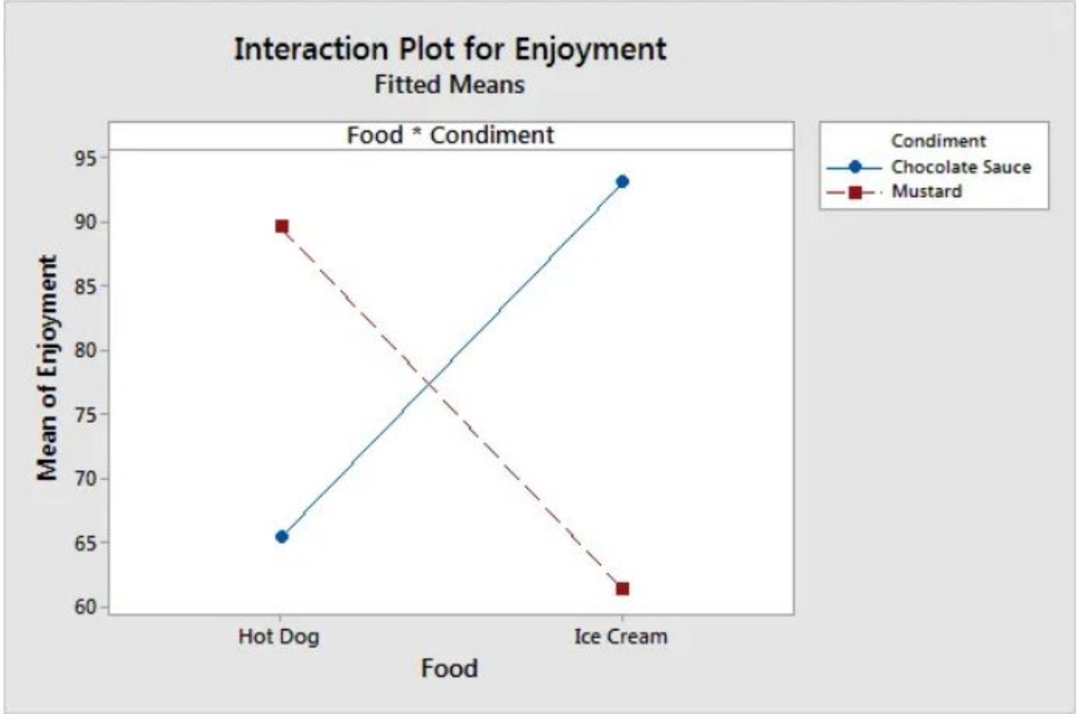
https://statisticsbyjim.com/regression/interaction-effects/

Visual aid: interaction plot

Plot predicted outcome

$\hat{y} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_{1*2} x_1 * x_2$

along y-axis



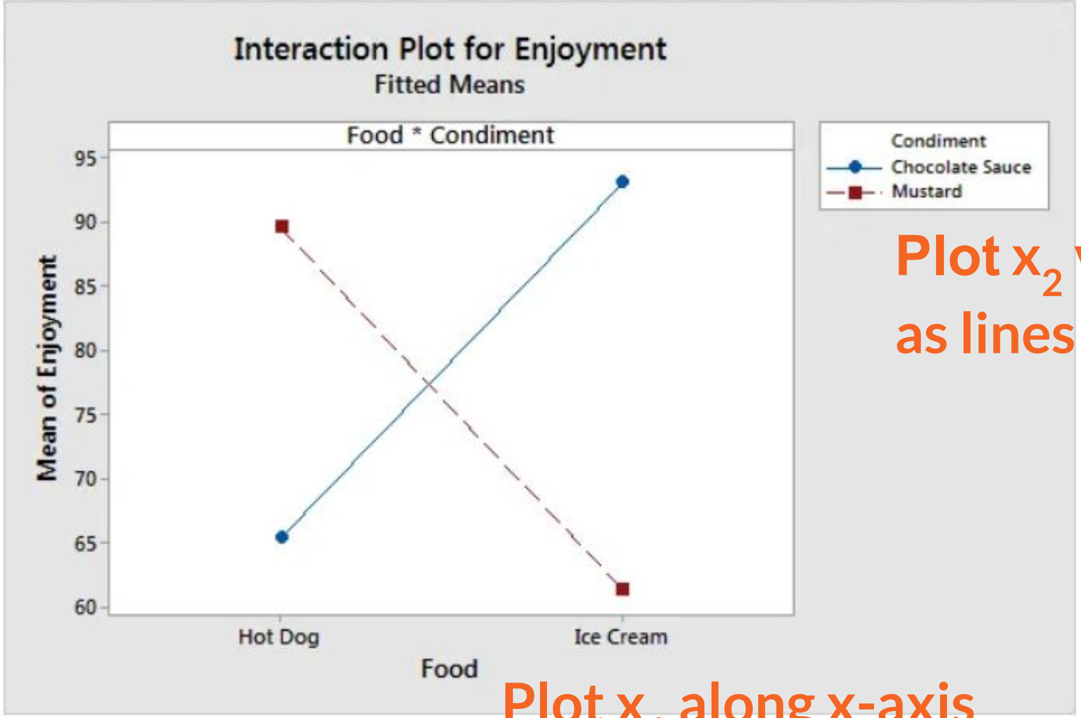
https://statisticsbyjim.com/regression/interaction-effects/

Visual aid: interaction plot

Plot predicted outcome

$\hat{y} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_{1*2} x_1 {}^* x_2$

along y-axis



Plot x_2 values as lines

Plot x_1 along x-axis

https://statisticsbyjim.com/regression/interaction-effects/

Multivar Regression: Interactions

- $y = 4.5 - 0.4x_1 - 0.5x_2 - 0.1x_1 * x_2$
- For categorical interactions, coefficients are **additive**

Demographic	x_1	x_2	Example	Value	Equation
White male	0	0	$\hat{y} = 4.5 - 0.4*0 - 0.5*0 - 0.1*0*0$	4.5	$\hat{y} = \alpha$
Non-white male	1	0	$\hat{y} = 4.5 - 0.4*1 - 0.5*0 - 0.1*1*0$	4.1	$\hat{y} = \alpha + \beta_1$
White non-male	0	1	$\hat{y} = 4.5 - 0.4*0 - 0.5*1 - 0.1*0*1$	4.0	$\hat{y} = \alpha + \beta_2$
Non-white non-male	1	1	$\hat{y} = 4.5 - 0.4*1 - 0.5*1 - 0.1*1*1$	3.5	$\hat{y} = \alpha + \beta_1 + \beta_2 + \beta_{1*2}$

Match the values!

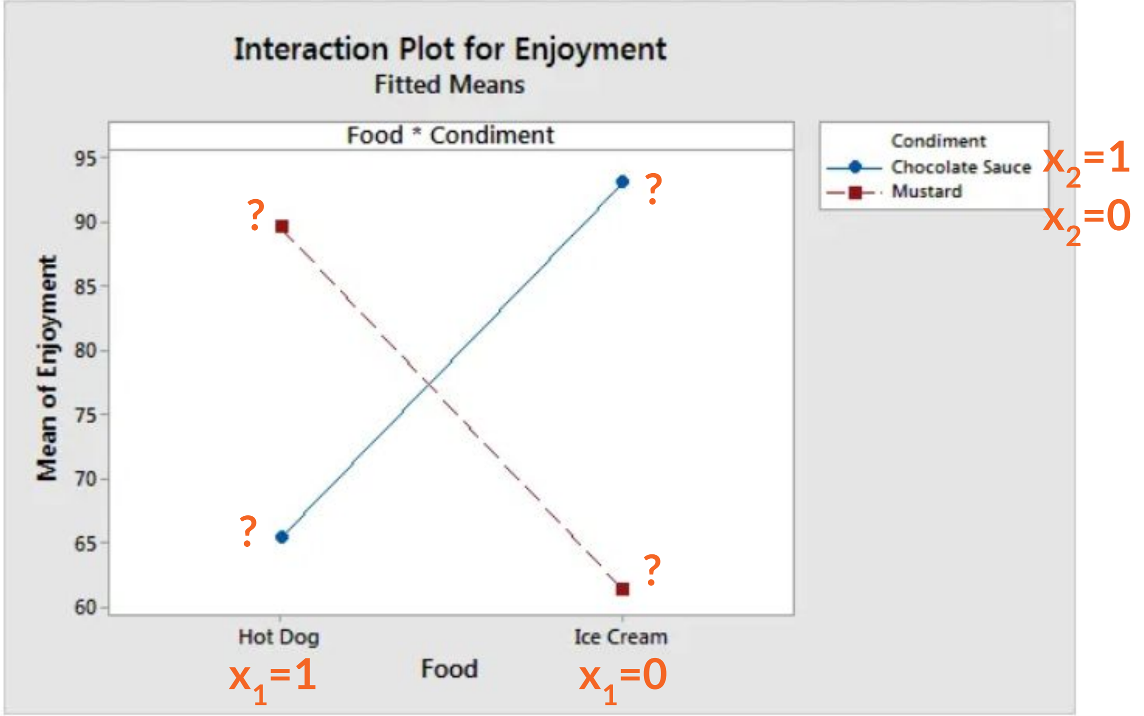
$y = 61.3 + 28.3x_1 + 31.7x_2 - 56.0x_1{}^*x_2$

$y^{hat} = \alpha$

$\alpha + \beta_1$

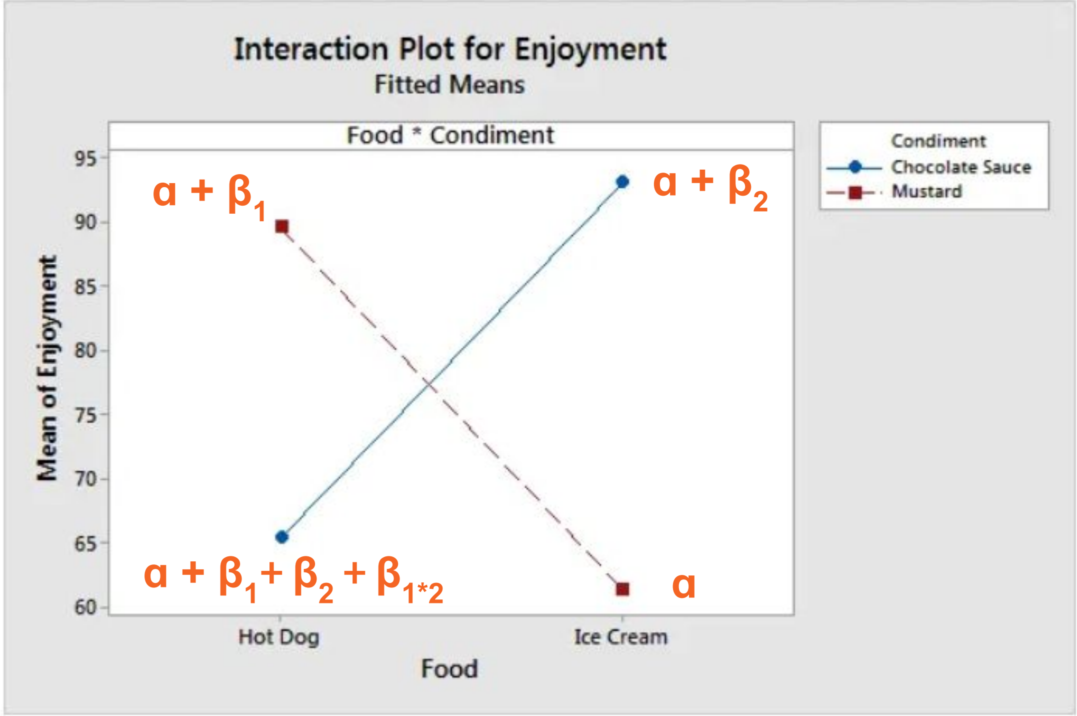
$\alpha + \beta_2$

$\alpha + \beta_1 + \beta_2 + \beta_{1*2}$



https://statisticsbyjim.com/regression/interaction-effects/

Visual aid: interaction plot

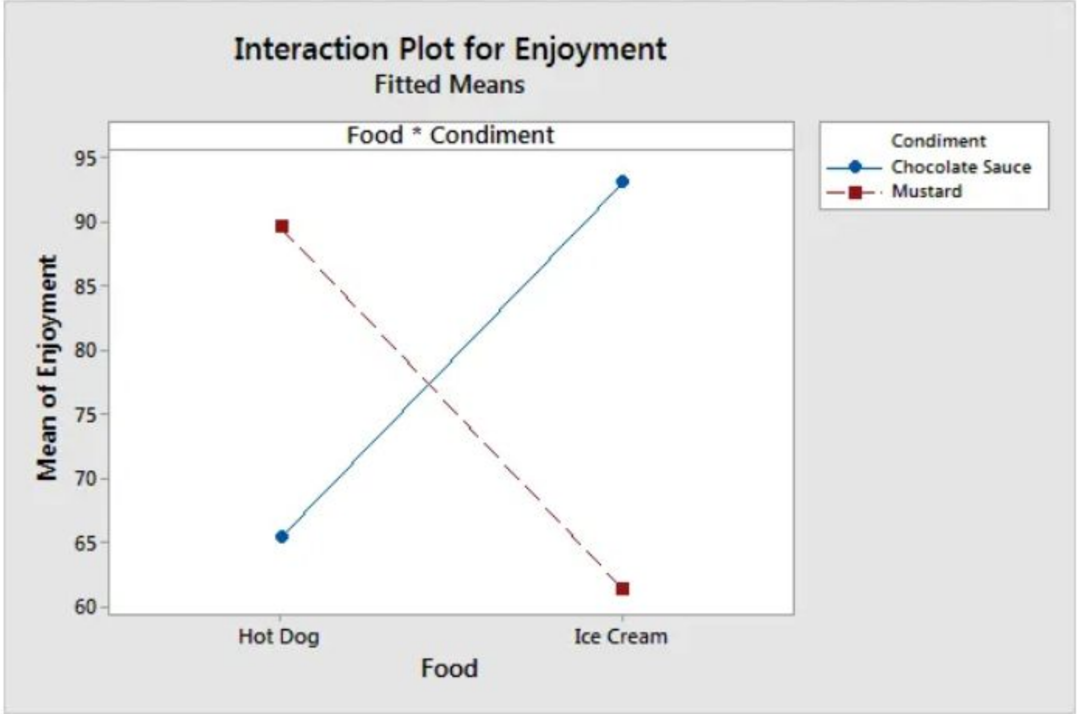


https://statisticsbyjim.com/regression/interaction-effects/

Visual aid: interaction plot

Parallel lines = no interaction effect

Crossing lines = has interaction effect



https://statisticsbyjim.com/regression/interaction-effects/

Multivar Regression: Interactions

- These examples were for one binary covariate against another binary covariate
- You can do interactions for any pair of covariate types (categorical x categorical, continuous x categorical, continuous x continuous)
- **How to interpret continuous interactions?** More complicated, **need to plot**

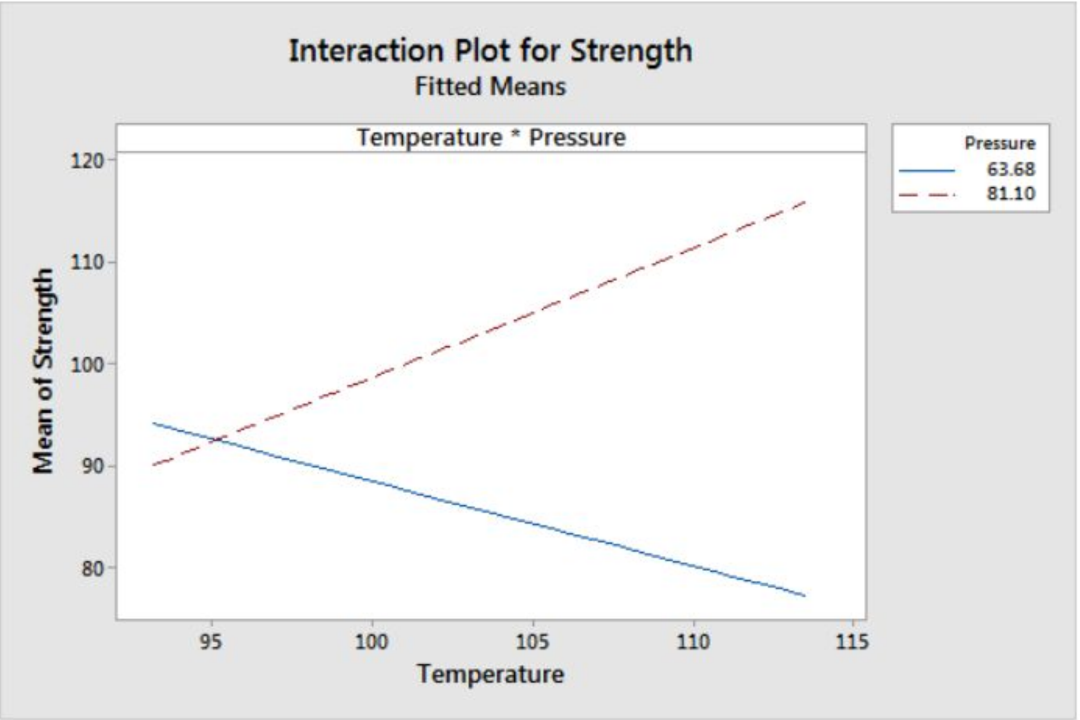
Multivar Regression: Continuous Interactions

y = Wind Strength

x_1 = Temperature

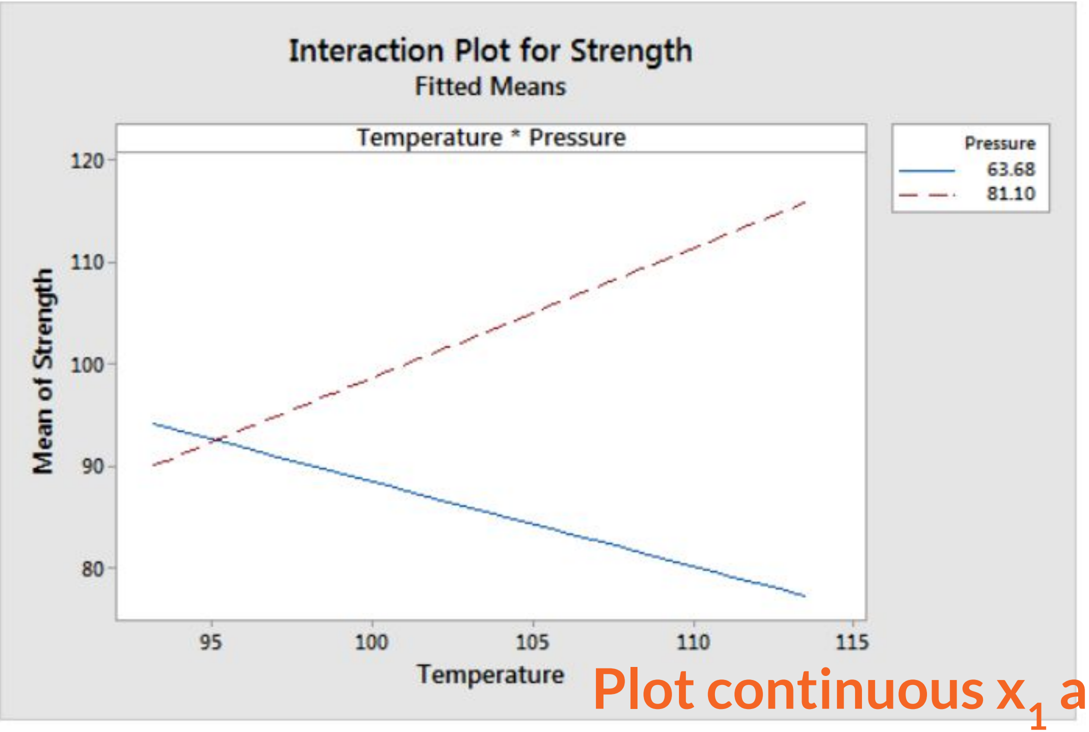
x_2 = Pressure

$y \sim x_1 + x_2 + x_1{}^{*}x_2$



https://statisticsbyjim.com/regression/interaction-effects/

Multivar Regression: Continuous Interactions



Plot continuous x_1 along x-axis

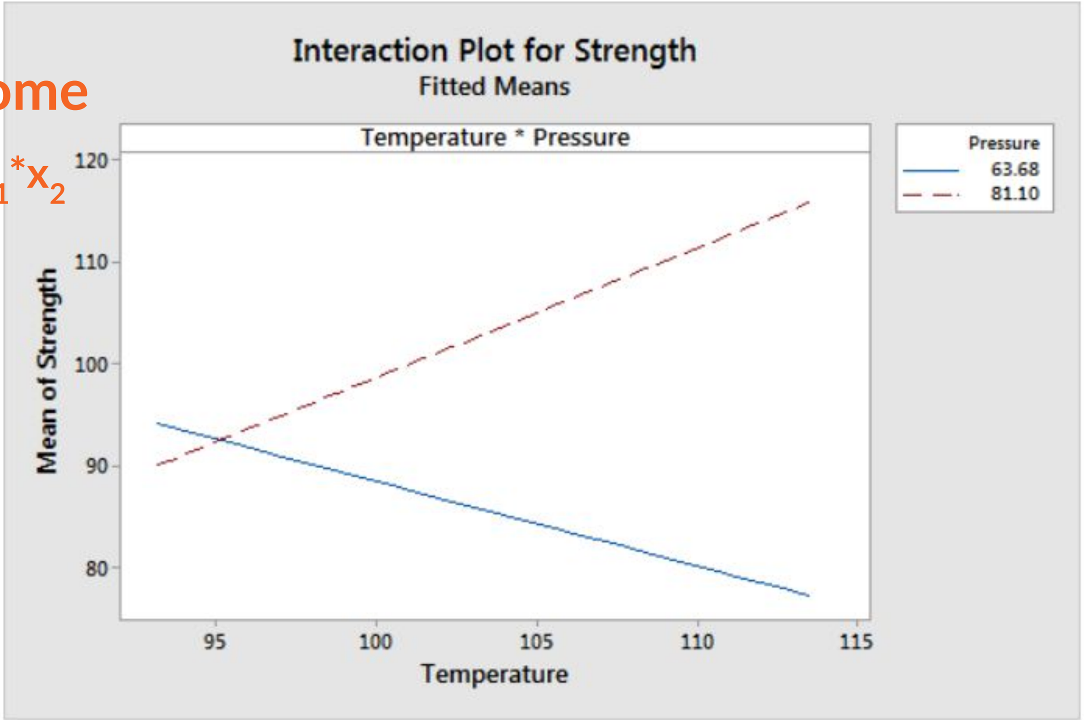
https://statisticsbyjim.com/regression/interaction-effects/

Multivar Regression: Continuous Interactions

Plot predicted outcome

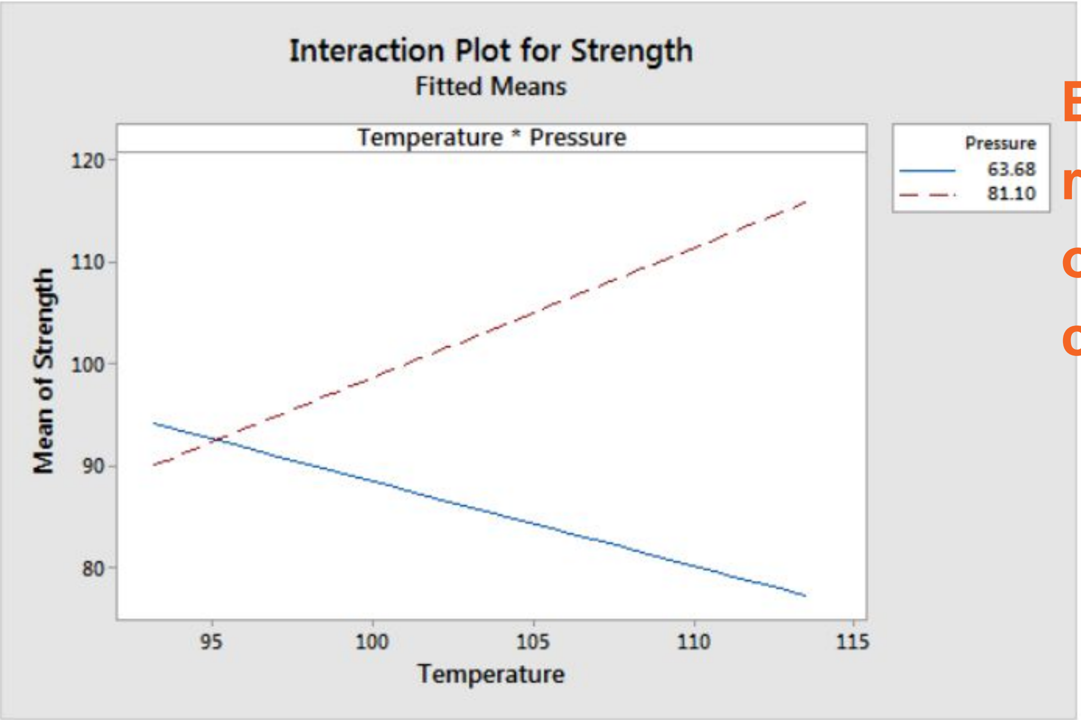
$\hat{y} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_{1*2} x_1 {}^* x_2$

along y-axis



https://statisticsbyjim.com/regression/interaction-effects/

Multivar Regression: Continuous Interactions

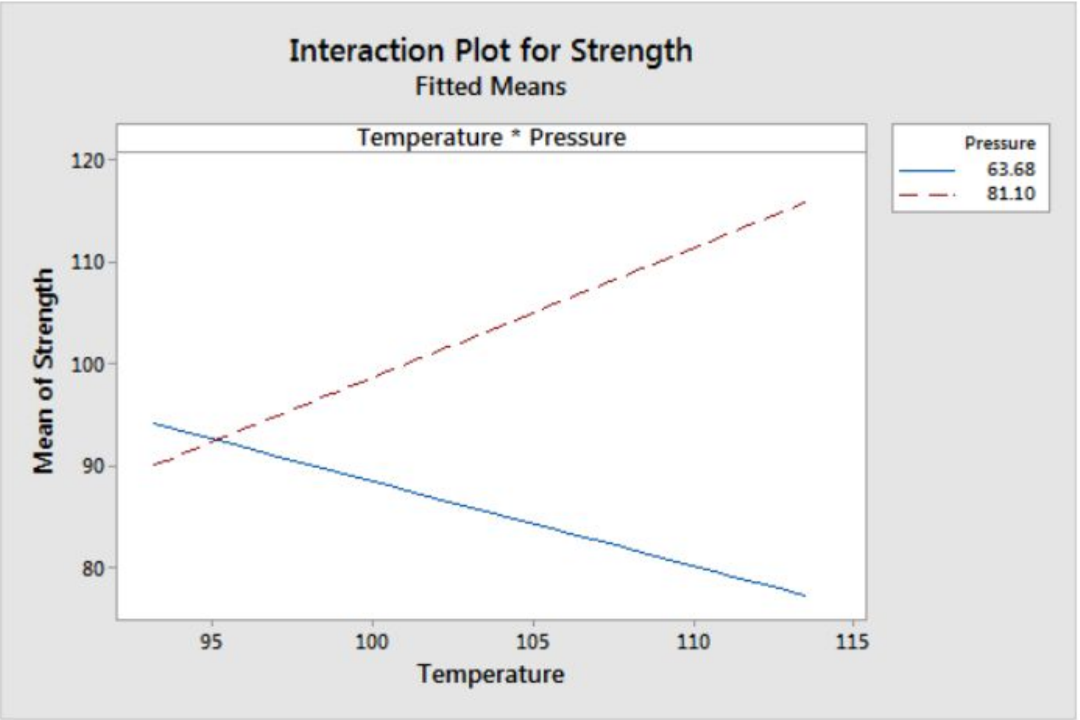


Each line represents one value of continuous x_2

https://statisticsbyjim.com/regression/interaction-effects/

Multivar Regression: Continuous Interactions

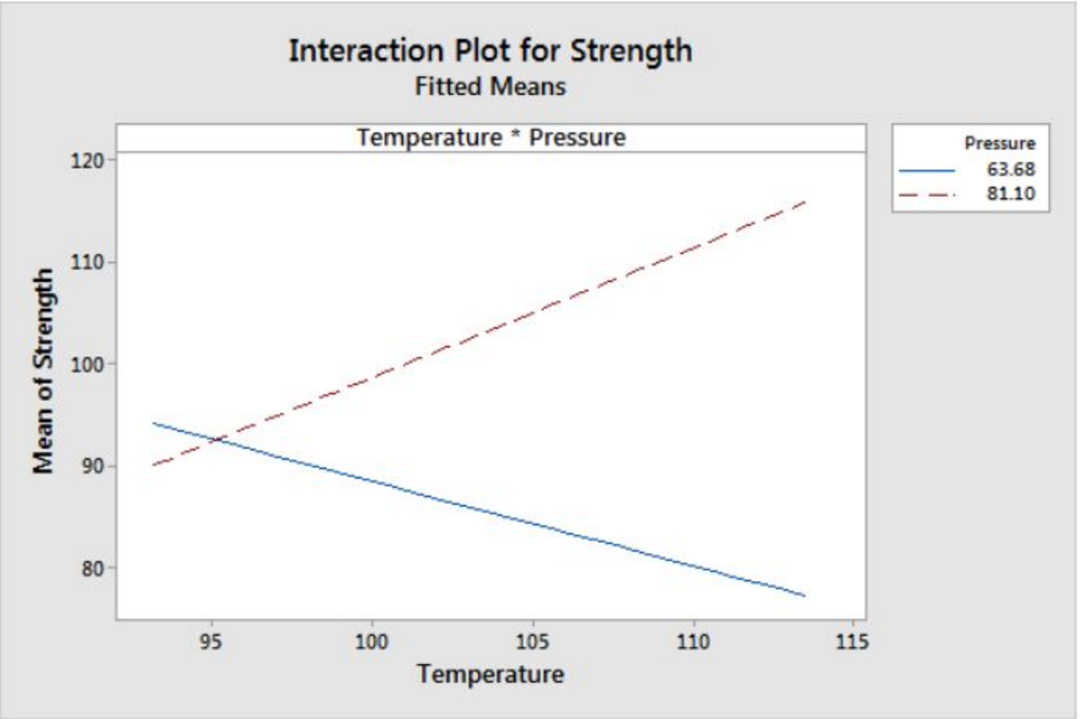
Is there an interaction effect here? →



https://statisticsbyjim.com/regression/interaction-effects/

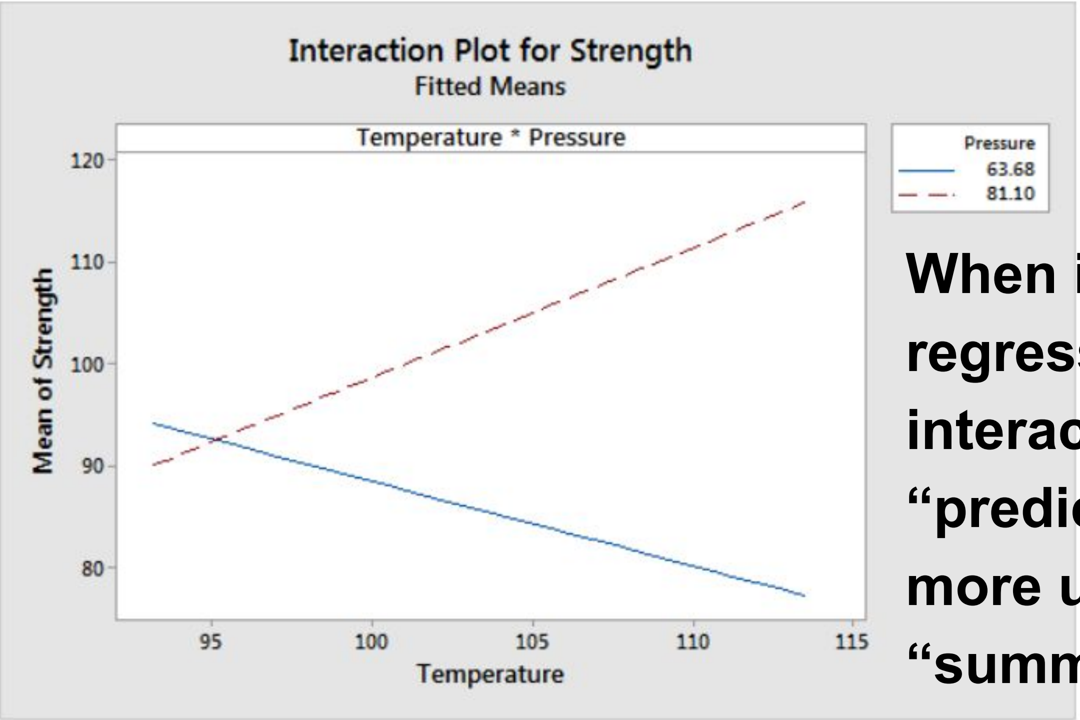
Multivar Regression: Continuous Interactions

Yes, the lines cross so an interaction effect exists. Make sure to include interactions in your regression model!



https://statisticsbyjim.com/regression/interaction-effects/

Multivar Regression: Continuous Interactions



When interpreting regressions with interactions, “prediction” often more useful than “summarization”

https://statisticsbyjim.com/regression/interaction-effects/

Multivariable Regression

- We've told you that Vitamin C increases the efficacy of Vitamin E when they're both ingredients in skincare products.
- **How would you model this?**
  - **y = customer satisfaction of product (*continuous*)**
  - **x_1 = ________ (*data type*), x_2 = ______ (*data type*)**
  - **Model: y ~ _____________**

Multivariable Regression

- We’ve told you that Vitamin C increases the efficacy of Vitamin E when they’re both ingredients in skincare products.

- **y = customer satisfaction of product (continuous),**
- **x_1 = Vitamin C in product (binary)**
- **x_2 = Vitamin E in product (binary)**
- **$y \sim x_1 + x_2 + x_1 * x_2$**

Multivariable Regression

- y = customer satisfaction of product (continuous),
- x_1 = Vitamin C in product (binary)
- x_2 = Vitamin E in product (binary)
- $y \sim x_1 + x_2 + x_1{*}x_2$
- $y = 6.0 + 54.0x_1 + 41.5x_2 - 6.0x_1{*}x_2$

Multivariable Regression

- y = customer satisfaction of product (continuous),
- x_1 = Vitamin C in product (binary)
- x_2 = Vitamin E in product (binary)
- $y \sim x_1 + x_2 + x_1{*}x_2$
- $y = 6.0 + 54.0x_1 + 41.5x_2 - 6.0x_1{*}x_2$
- When is customer satisfaction highest?
- When is customer satisfaction lowest?

Multivariable Regression

- y = customer satisfaction of product (continuous),
- x_1 = Vitamin C in product (binary)
- x_2 = Vitamin E in product (binary)
- $y = 6.0 + 54.0x_1 + 41.5x_2 - 6.0x_1{*}x_2$
- When there is both vitamin C and vitamin E in the product, we predict customer satisfaction to be the highest, at 6.0 + 54.0*1 + 41.5 *1- 6.0*1 = 95.5.
- When there is no vitamin C and no vitamin E in the product, we predict customer satisfaction is lowest, at 6.0 + 54.0*0 + 41.5*0 - 6.0*0 = 6.
- When there is only vitamin C in the product, we predict customer satisfaction is 6.0 + 54.0*1 + 41.5*0 - 6.0*0 = 60. When there is only vitamin E in the product, we predict customer satisfaction is 6.0 + 54.0*0 + 41.5*1 - 6.0*0 = 47.5.

Multivariable Regression

- y = customer satisfaction of product (continuous),
- x_1 = Vitamin C in product (binary)
- x_2 = Vitamin E in product (binary)
- $y \sim x_1 + x_2 + x_1{*}x_2$
- $y = 6.0 + 54.0x_1 + 41.5x_2 \boxed{- 6.0} x_1{*}x_2$

This coefficient is negative, but it doesn't mean the interaction gives us a lower ŷ!

Multivariable Regression

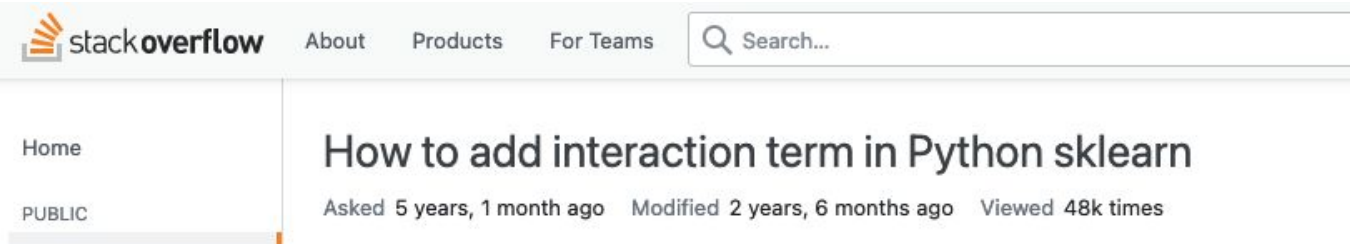
- y = customer satisfaction of product (continuous),
- x_1 = Vitamin C in product (binary)
- x_2 = Vitamin E in product (binary)
- $y \sim x_1 + x_2 + x_1 * x_2$
- $y = 6.0 + 54.0x_1 + 41.5x_2 - 6.0x_1 * x_2$

This is why we don't just read off the coefficient to "summarize" interactions!

How to run regressions with interactions in Python?

How to run regressions with interactions in Python?

- Google it!



https://stackoverflow.com/questions/45828964/how-to-add-interaction-term-in-python-sklearn

How to run regressions with interactions in Python?

- A few different ways to make the interaction:
    - Manually multiply two columns in your dataframe and include that third column in your X input
    - Use PolynomialFeatures() from sklearn
    - Use another package (like *patsy*) to use ~ format
- Then, run LinearRegression().fit(X,y) as usual

Takeaways

- Interactions ≈ Intersectionality
- Use "interaction plots" to determine whether you need an interaction term in your regression model
- Interpret regressions with interactions by plugging in different values of the x's and ***predicting outcomes***. Do not rely on reading the coefficients to summarize the effects!

1 min break!

https://anyi-guo.medium.com/correlation-pearson-vs-spearman-c15e581c12ce

Transformations

- We've already talked about a few transformations:
  - **Log (ln)**, **sqrt**, etc. to deal with heteroskedasticity, big numbers, funny residual plots, etc.
  - **Binary**, e.g. 1 categorical → multiple dummy variables

Transformations

- We’ve already talked about a few transformations:
  - **Log (ln)**, **sqrt**, etc. to deal with heteroskedasticity, big numbers, funny residual plots, etc.
  - **Binary**, e.g. 1 categorical → multiple dummy variables
- New transformation: **rank transform**

Rank transformations

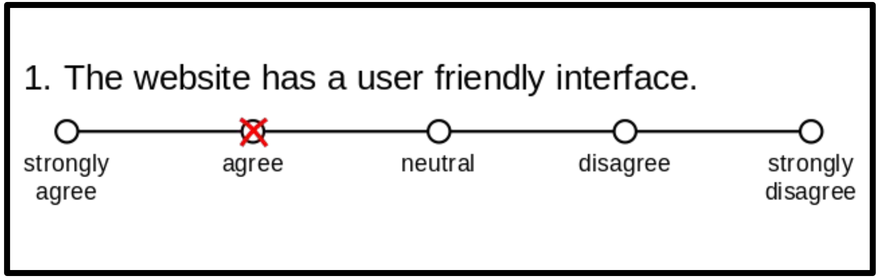
- What if you don't care about the **value** of a set of predictions, but rather just the **ranked order** of them?
  - **Any examples?**

Rank transformations

- What if you don't care about the **value** of a set of predictions, but rather just the **ranked order** of them?
    - Google search results
    - Netflix movie recommendations
    - Competitions
    - Survey responses on a Likert scale

Likert Scales

(ranking not always necessary, but Likert Scales is an important vocab word for many DS projects!)

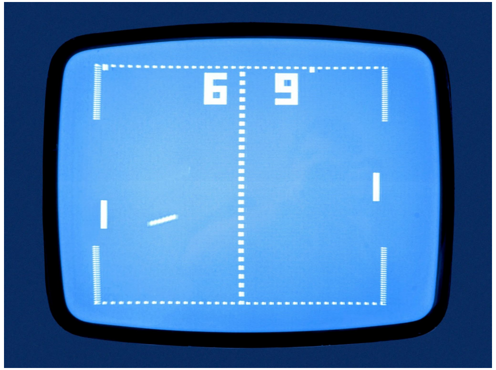


https://www.questionpro.com/article/likert-scale-survey-questions.html , https://www.wikidata.org/wiki/Q617473

Rank transformations

- What if you don't care about the **value** of a set of predictions, but rather just the **ranked order** of them?
    - Google search results
    - Netflix movie recommendations
    - Competitions
    - Survey responses on a Likert scale

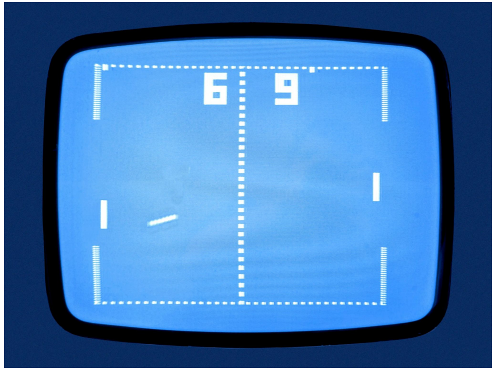
Numeric values → Ranks



Player	# Pong Wins	Pong Rank
A	100	1
B	4	?
C	0	?
D	25	?
E	1	?

https://www.wired.com/story/inside-story-of-pong-excerpt/

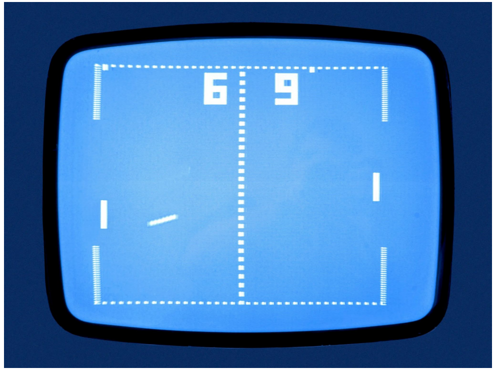
Numeric values → Ranks



Player	# Pong Wins	Pong Rank
A	100	1
B	4	3
C	0	5
D	25	2
E	1	4

https://www.wired.com/story/inside-story-of-pong-excerpt/

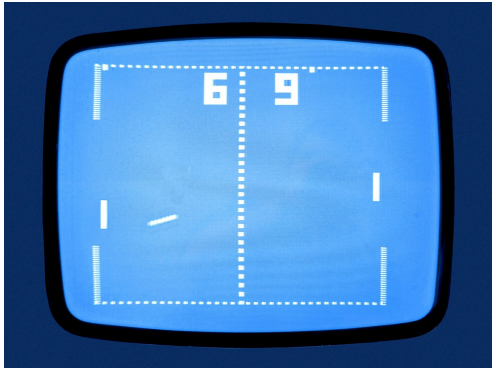
Numeric values → Ranks



Player	# Pong Wins	Pong Rank
A	100	1
B	4	3
C	0	?
D	25	2
E	0	?

https://www.wired.com/story/inside-story-of-pong-excerpt/

Numeric values → Ranks



Player	# Pong Wins	Pong Rank
A	100	1
B	4	3
C	0	**4? 4.5? 5?**
D	25	2
E	0	**4? 4.5? 5?**

https://www.wired.com/story/inside-story-of-pong-excerpt/

Numeric values → Ranks

- Use a function that lets you define how to break ties

pandas.DataFrame.rank

```
DataFrame.rank(axis=0, method='average',
numeric_only=_NoDefault.no_default, na_option='keep', ascending=True,
pct=False)
```
[source]

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.rank.html

Numeric values → Ranks

- Use a function that lets you define how to break ties

pandas.DataFrame.rank

```
DataFrame.rank(axis=0, method='average',
numeric_only=_NoDefault.no_default, na_option='keep', ascending=True,
pct=False)
```
[source]

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.rank.html

Numeric values → Ranks

```
>>> df
    Animal  Number_legs
0      cat          4.0
1  penguin          2.0
2      dog          4.0
3   spider          8.0
4    snake          NaN
```

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.rank.html

Numeric values → Ranks

```
>>> df['default_rank'] = df['Number_legs'].rank()
```

```
>>> df
    Animal  Number_legs  default_rank
0      cat          4.0           2.5
1  penguin          2.0           1.0
2      dog          4.0           2.5
3   spider          8.0           4.0
4    snake          NaN           NaN
```

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.rank.html

Numeric values → Ranks

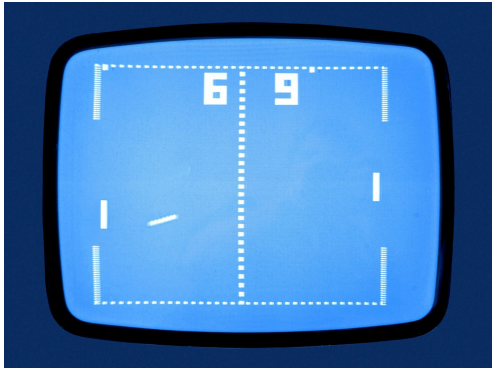
```
>>> df['default_rank'] = df['Number_legs'].rank()
>>> df['max_rank'] = df['Number_legs'].rank(method='max')
>>> df
    Animal  Number_legs  default_rank  max_rank
0      cat          4.0           2.5       3.0
1  penguin          2.0           1.0       1.0
2      dog          4.0           2.5       3.0
3   spider          8.0           4.0       4.0
4    snake          NaN           NaN       NaN
```

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.rank.html

Rank transformations

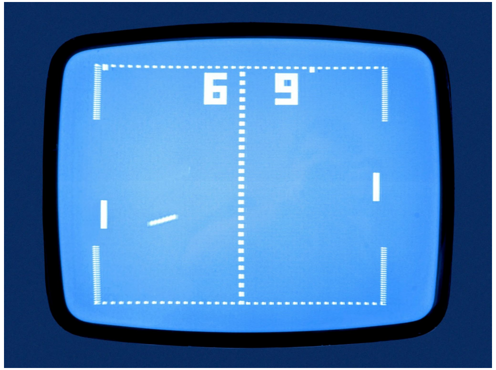
- We can use pandas to generate columns that **rank** another column's data
  - We can even do this in a few different ways considering ties
- What if we now have **multiple** ranks for each row?
  - What if we want to **compare** those ranks?

Are good pong players are also good Beirut players?



Player	Pong Rank	Beirut Rank
A	1	2
B	3	4
C	5	5
D	2	3
E	4	1

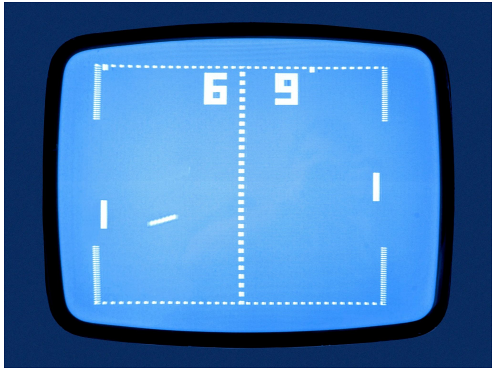
Are good pong players are also good Beirut players?



Hypothesis: quick reflexes are a transferable skill, so I expect that good pong players are also good Beirut players

Player	Pong Rank	Beirut Rank
A	1	2
B	3	4
C	5	5
D	2	3
E	4	1

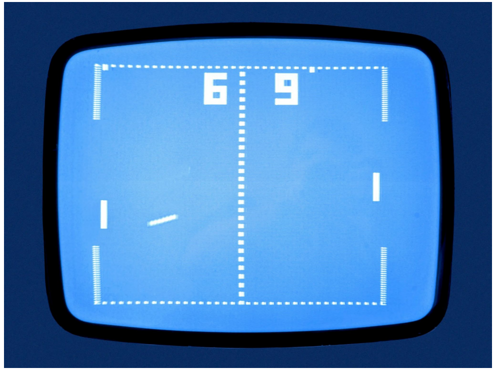
Are good pong players are also good Beirut players?



How do we measure this hypothesis using rank data?

Player	Pong Rank	Beirut Rank
A	1	2
B	3	4
C	5	5
D	2	3
E	4	1

How do we tell if two variables are similar?

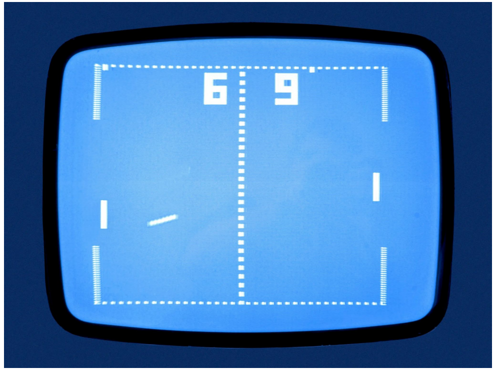


Player	Pong Rank	Beirut Rank
A	1	2
B	3	4
C	5	5
D	2	3
E	4	1

Covariance divided by the product of standard deviations is correlation

$$\frac{\mathrm{cov}(X, Y)}{\sigma_x \sigma_y}$$

How do we tell if two rank variables are similar?



Player	Pong Rank	Beirut Rank
A	1	2
B	3	4
C	5	5
D	2	3
E	4	1

Spearman vs Pearson Correlations

- We’ve already learned about “Pearson correlation”
  - It’s what we commonly call just “correlation”
  - Computed based on **values**
  - Used to understand **linear relationships**

Pearson himself

- We've already learned about "Pearson correlation"
  - It's what we commonly call just "correlation"
  - Computed based on **values**
  - Used to understand **linear relationships**

https://en.wikipedia.org › wiki › Karl_Pearson

Karl Pearson - Wikipedia

Pearson was also a proponent of social Darwinism, **eugenics** and scientific racism. **Pearson** was a protégé and biographer of Sir Francis Galton.

Spearman vs Pearson Correlations

- We've already learned about "Pearson correlation"
  - It's what we commonly call just "correlation"
  - Computed based on **values**
  - Used to understand **linear relationships**

- **New type of correlation**: **"Spearman correlation"**
  - Computed based on **ranks**
  - Used to understand **monotonic relationships**

Spearman himself

https://en.wikipedia.org › wiki › Spearman's_hypothesis

Spearman's hypothesis - Wikipedia

Claims of validity of **Spearman's** hypothesis have been criticized on methodological grounds. Such claims have been used to support scientific **racism**.

Description · Related hypotheses · Group differences

- New type of correlation: "Spearman correlation"
    - Computed based on **ranks**
    - Used to understand **monotonic relationships**

Spearman Correlations

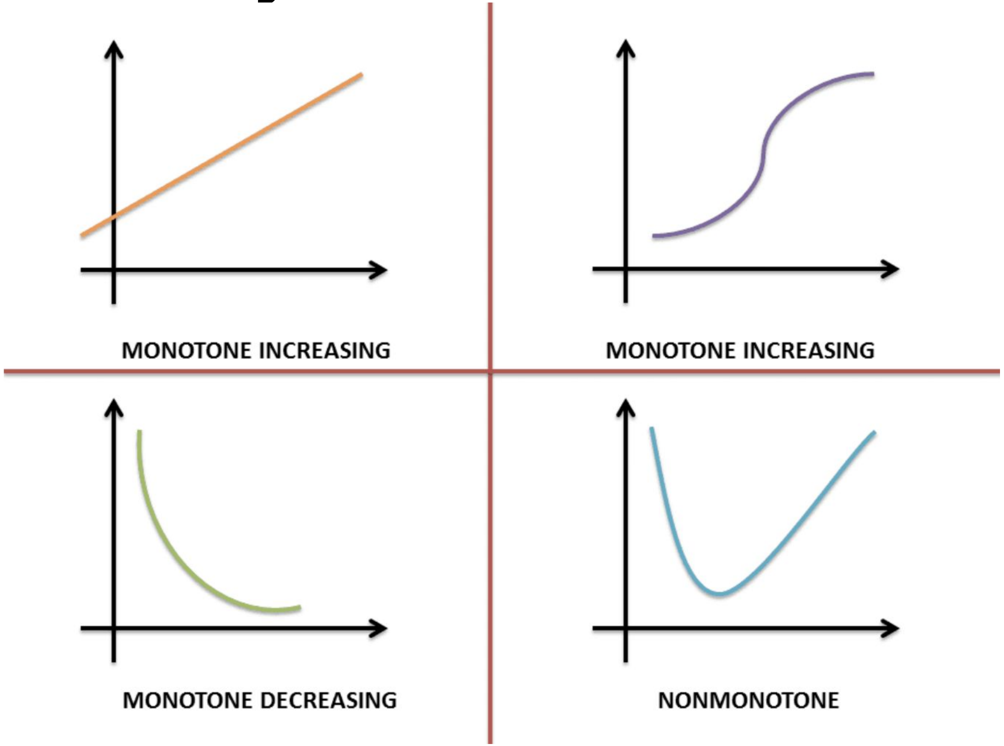
- New type of correlation: "Spearman correlation"
    - Computed based on **ranks**
    - Used to understand **monotonic relationships**

Monotonicity

- To have a monotonic relationship, one of the following **must** be true:
  - As the value of one variable **increases**, the other variable value **increases**
  - As the value of one variable **increases**, the other variable value **decreases**

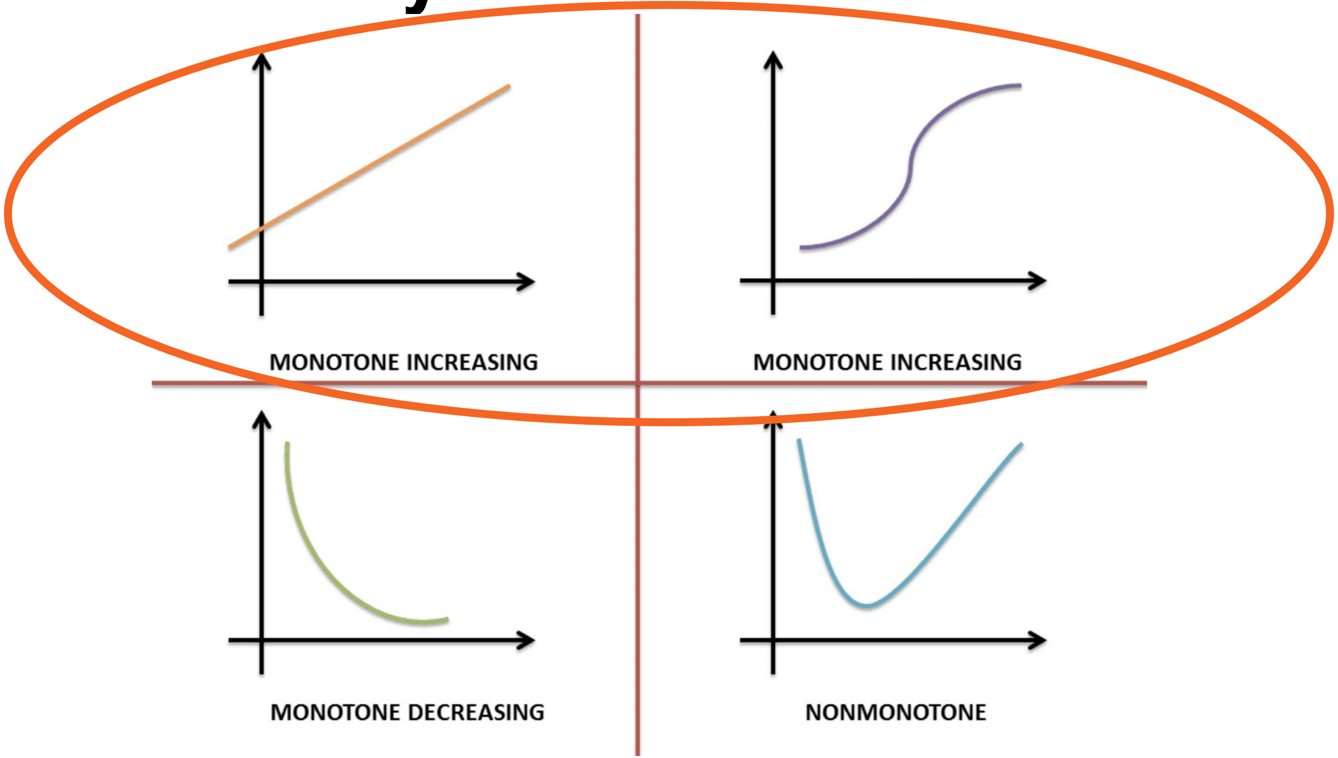
https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

Monotonicity

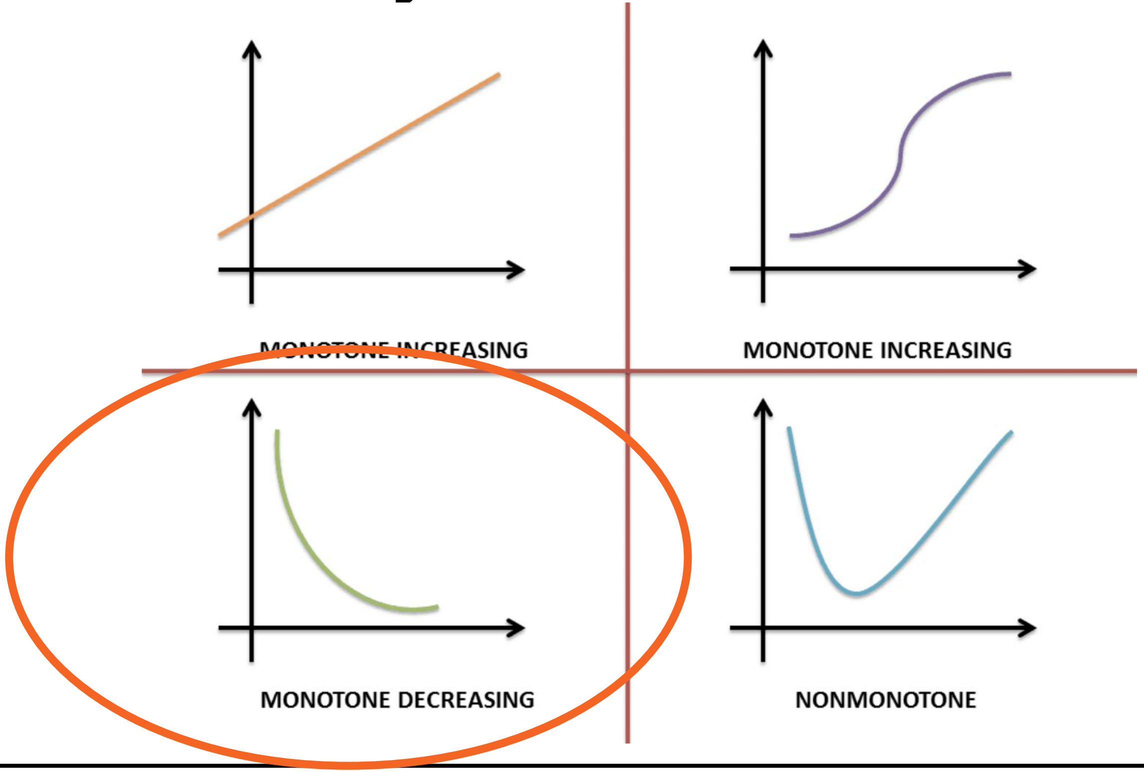


https://discover.hubpages.com/education/Differences-Between-Spearman-Rank-Correlation-and-Pearson-Correlation

Monotonicity

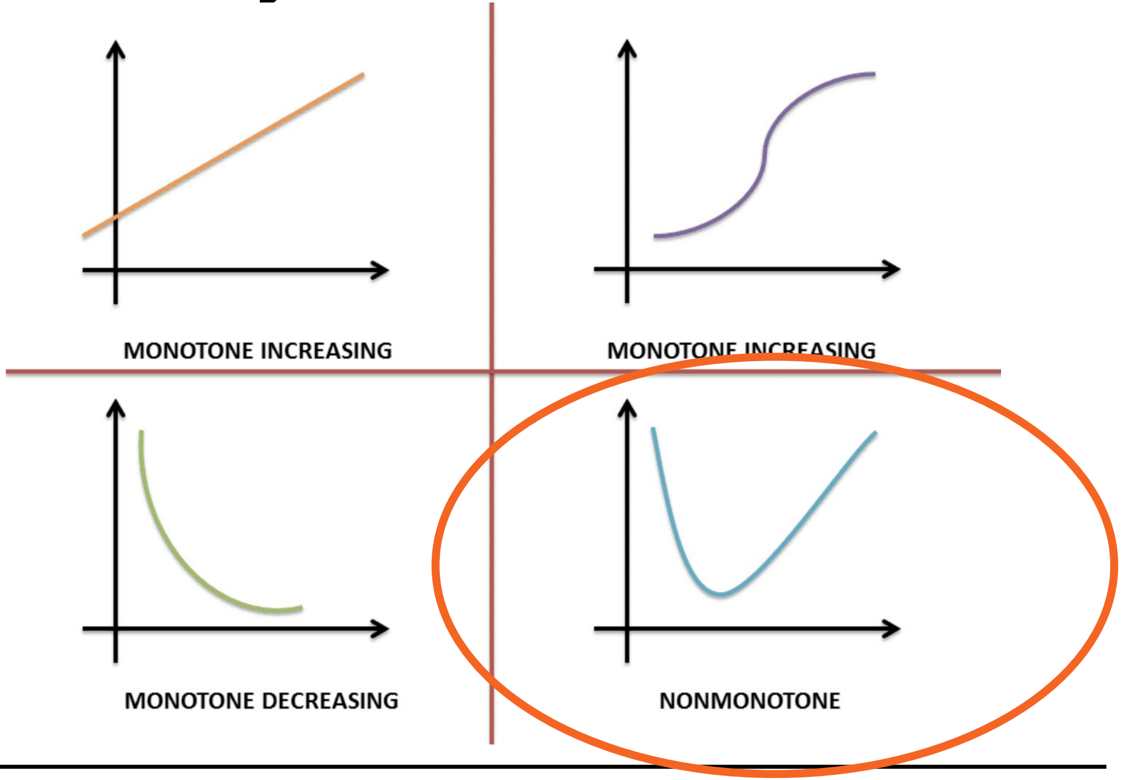


Monotonicity



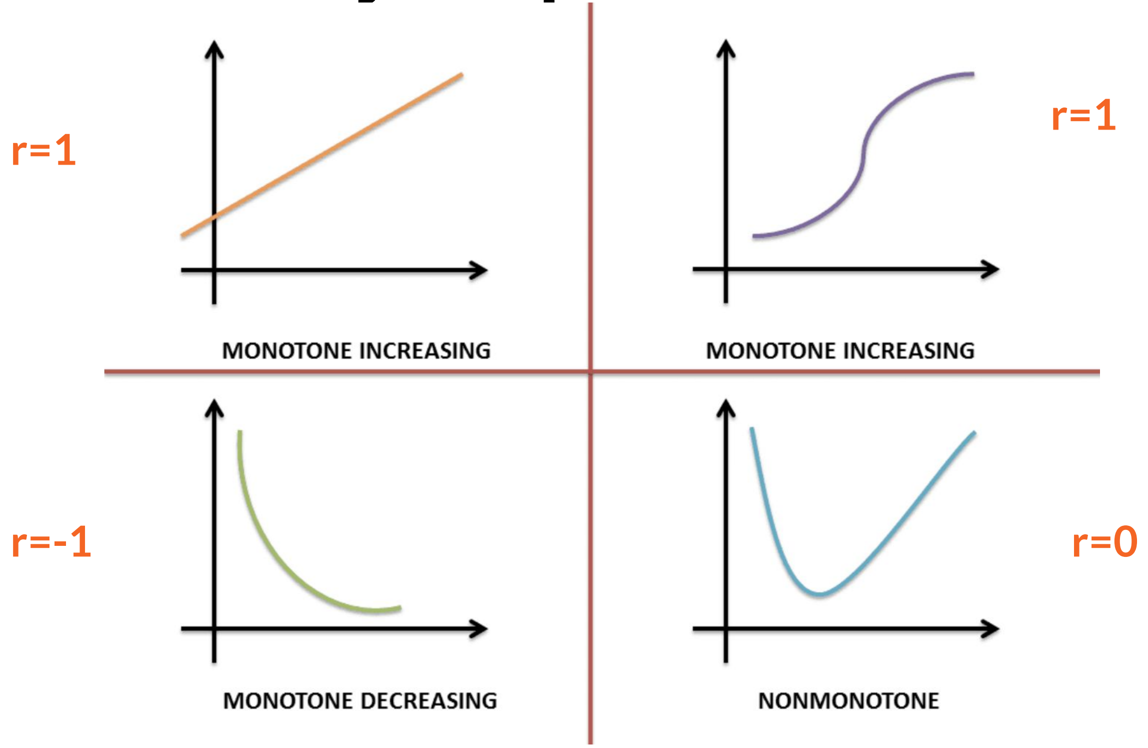
https://discover.hubpages.com/education/Differences-Between-Spearman-Rank-Correlation-and-Pearson-Correlation

Monotonicity



https://discover.hubpages.com/education/Differences-Between-Spearman-Rank-Correlation-and-Pearson-Correlation

Monotonicity & Spearman corr

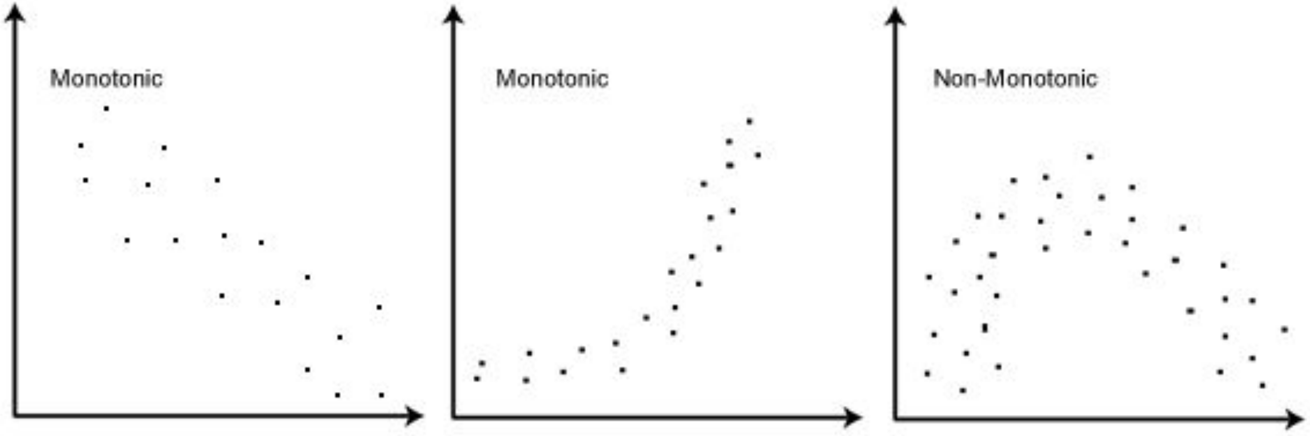


https://discover.hubpages.com/education/Differences-Between-Spearman-Rank-Correlation-and-Pearson-Correlation

Each dot
represents a player

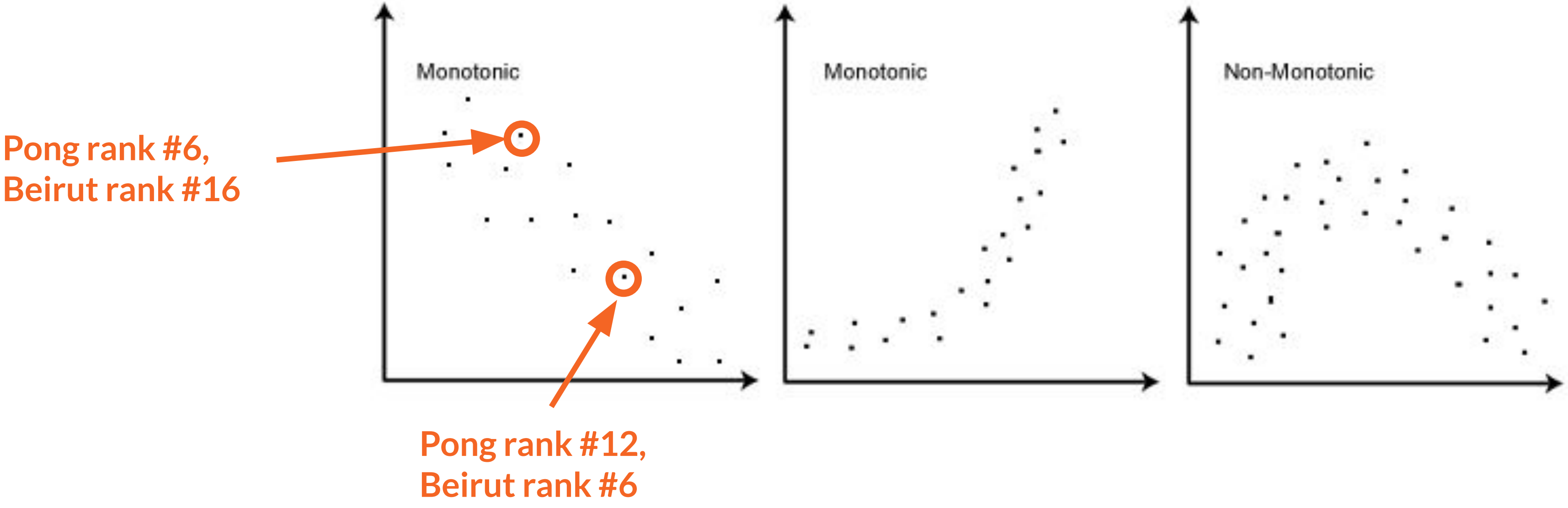
x-axis: pong rank
y-axis: beirut rank

Monotonicity



https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

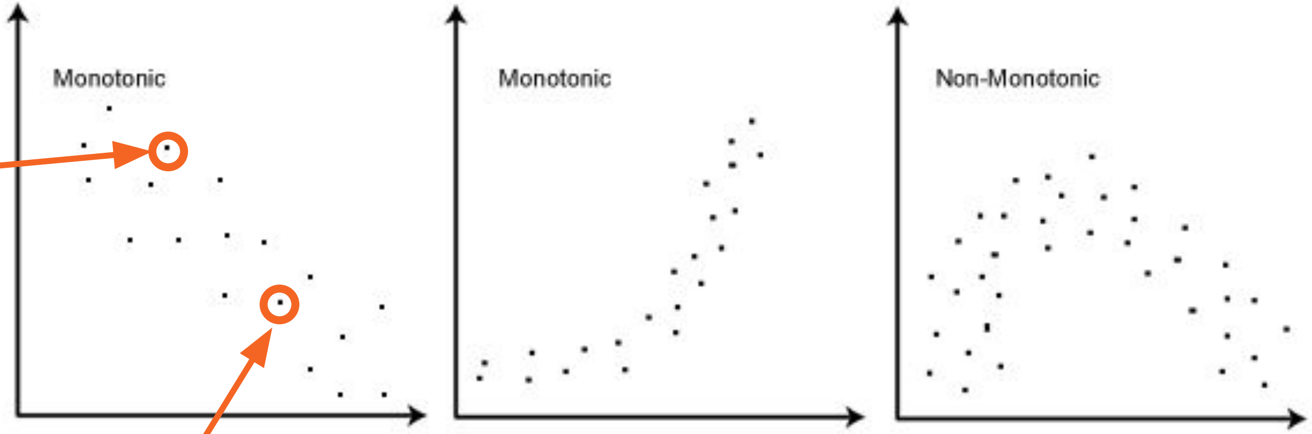
Monotonicity



https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

Monotonicity

Pong rank #6,
Beirut rank #16

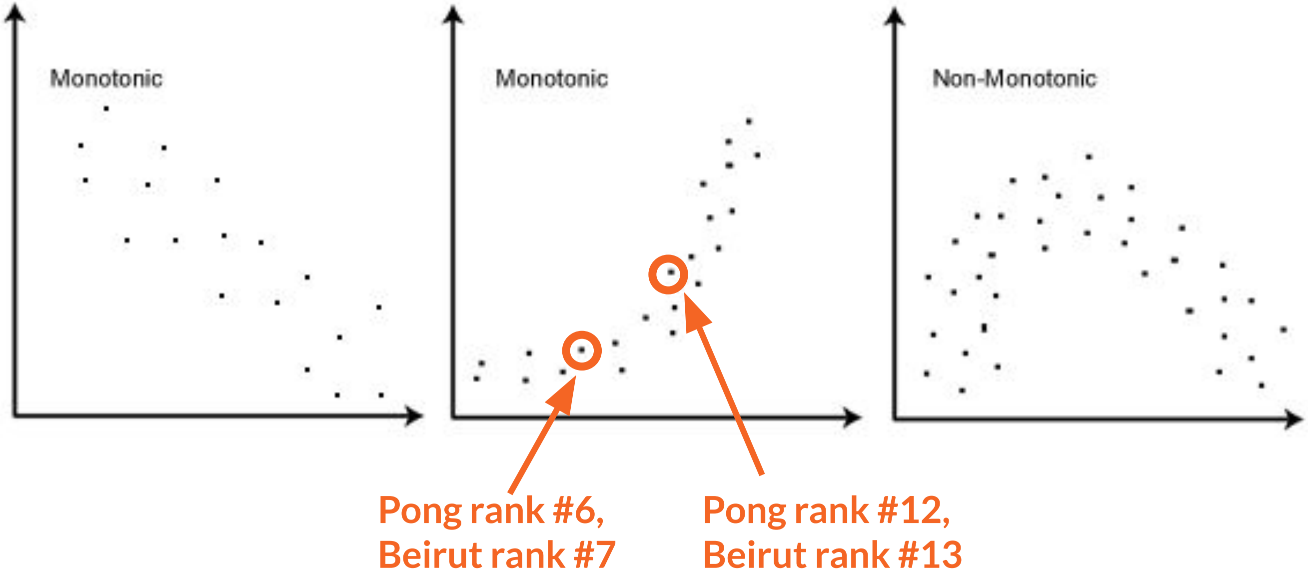


Pong rank #12,
Beirut rank #6

We observe a [close to] monotone decrease in Beirut rank (as a function of Pong rank)

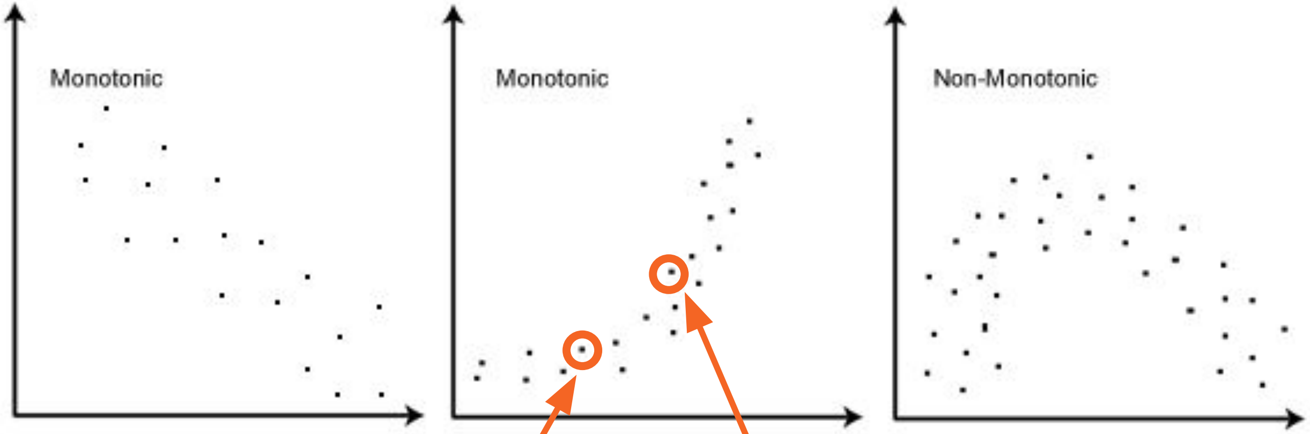
https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

Monotonicity



https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

Monotonicity



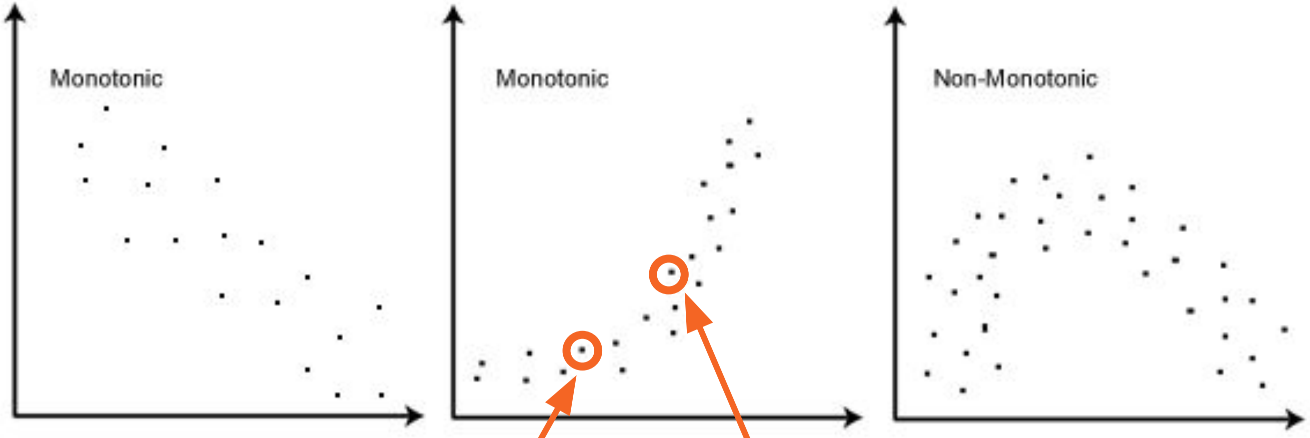
We observe a [close to] monotone _______ in Beirut rank (as a function of Pong rank)

Pong rank #6,
Beirut rank #7

Pong rank #12,
Beirut rank #13

https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

Monotonicity



We observe a [close to] monotone increase in Beirut rank (as a function of Pong rank)

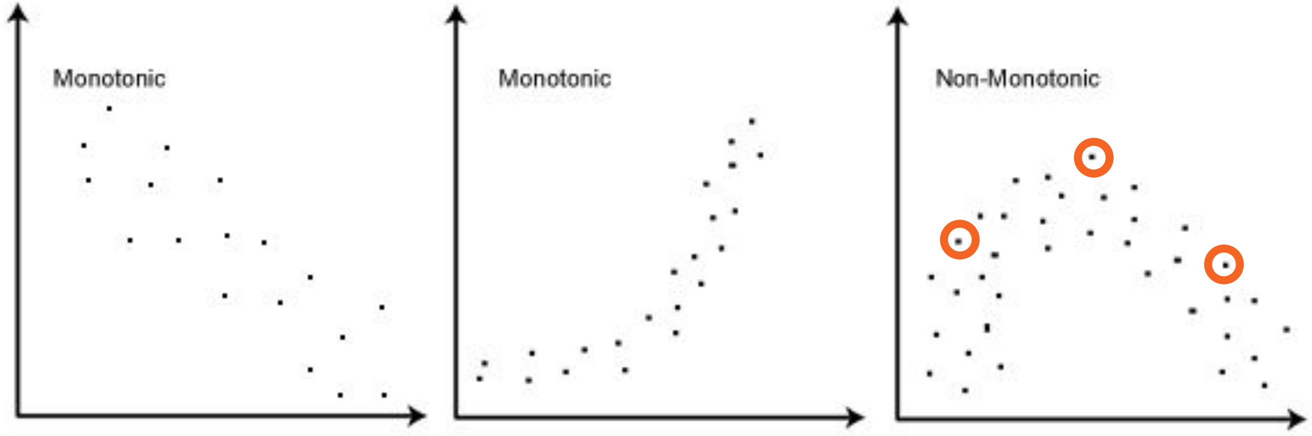
Pong rank #6,
Beirut rank #7

Pong rank #12,
Beirut rank #13

https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

Monotonicity

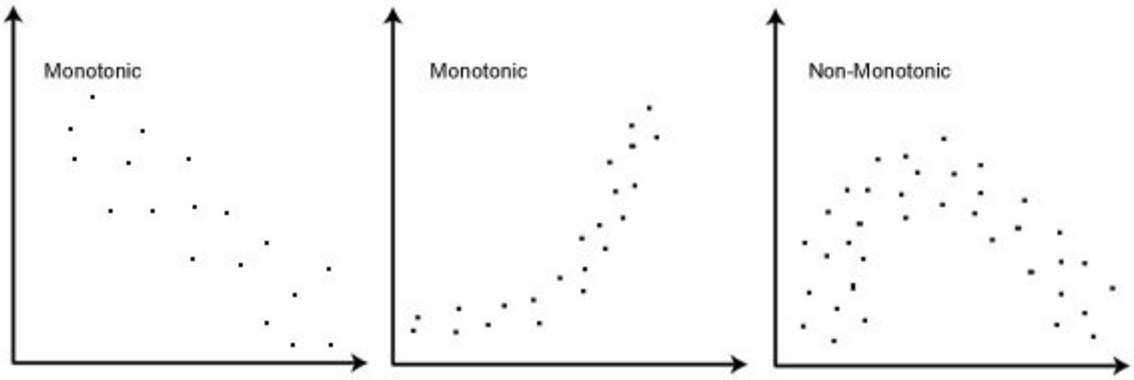
These pong and Beirut ranks clearly have *some* relationship, but it's not monotonic!



https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

Monotonicity & Spearman corr

r≈-1 r≈1 r≈0



https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

Guess the Spearman correlation

Player	Pong Rank	Beirut Rank
A	1	2
B	3	4
C	5	5
D	2	3
E	4	1

Guess: do you think Pong rank and Beirut rank will have a Spearman correlation that is..?

- positive/negative
- high/low

Spearman correlation (algo)

1. Get list of rankings for each variable
2. Find the difference in ranks for each row (d)
3. Calculate the sum of d^2 over all rows
4. Calculate r = $1 - \left(\frac{6 \Sigma d^2}{n^3 - n} \right)$

Calculate Spearman corr

Player	Pong Rank	Beirut Rank
A	1	2
B	3	4
C	5	5
D	2	3
E	4	1

1. Get list of rankings for each variable
2. Find the difference in ranks for each row (d)
3. Calculate the sum of d^2 over all rows
4. Calculate r = $1 - \left(\frac{6 \Sigma d^2}{n^3 - n} \right)$

Spearman correlation (example)

Player	Pong Rank	Beirut Rank	d
A	1	2	1
B	3	4	1
C	5	5	0
D	2	3	1
E	4	1	-3

$d^2 = 1^2 + 1^2 + 0^2 + 1^2 + (-3)^2 = 12$

$r = 1 - (6*12 / (5^3 - 5))$
$= 1 - 72/120 = 0.4$

Spearman correlation (example)

Player	Pong Rank	Beirut Rank	d
A	1	2	1
B	3	4	1
C	5	5	0
D	2	3	1
E	4	1	-3

$$d^2 = 1^2 + 1^2 + 0^2 + 1^2 + (-3)^2 = 12$$

$$r = 1 - (6*12 / (5^3 - 5))$$
$$= 1 - 72/120 = 0.4$$

Is this close to your guess?

Spearman correlation (example)

Player	Pong Rank	Beirut Rank	d
A	1	2	1
B	3	4	1
C	5	5	0
D	2	3	1
E	4	1	-3

$d^2 = 1^2 + 1^2 + 0^2 + 1^2 + (-3)^2 = 12$

$r = 1 - (6*12 / (5^3 - 5))$
$= 1 - 72/120 = 0.4$

As expected, kind of a middle-ish number – not *no* correlation, but not strong either

Spearman correlation (Python)

```python
from scipy import stats
stats.spearmanr([1,3,5,2,4],[2,4,5,3,1])
```

✓ 0.6s Python

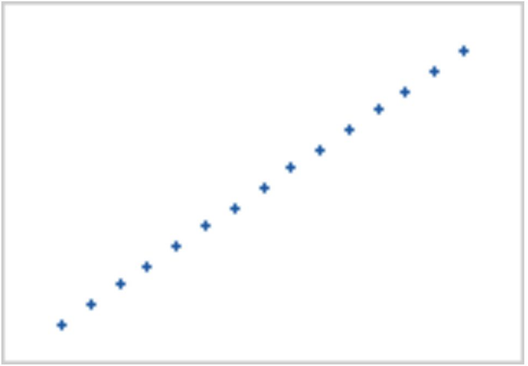
```
SpearmanrResult(correlation=0.39999999999999997,
pvalue=0.5046315754686911)
```

Spearman vs Pearson: Similarities

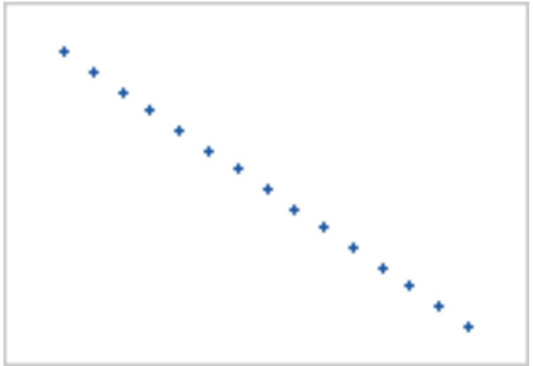
- Both correlations yield coefficients between -1 and 1

- Higher coefficient magnitude → stronger relationship between variables

How exactly does Spearman differ from Pearson correlation?



Pearson = +1, Spearman = +1

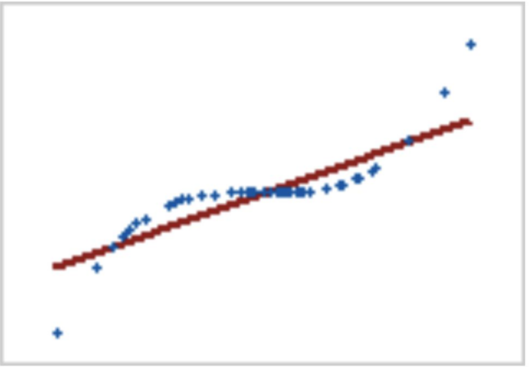


Pearson = −1, Spearman = −1

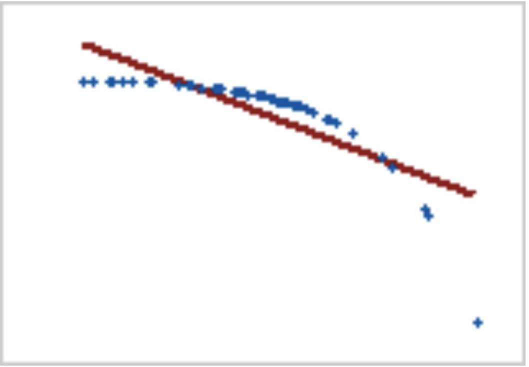
https://support.minitab.com/en-us/minitab/21/help-and-how-to/statistics/basic-statistics/supporting-topics/correlation-and-covariance/a-comparison-of-the-pearson-and-spearman-correlation-methods/

How exactly does Spearman differ from Pearson correlation?

Blue dots are the data; red line is linear regression



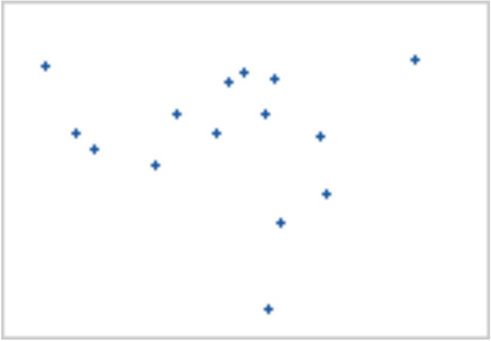
Pearson = +0.851, Spearman = +1



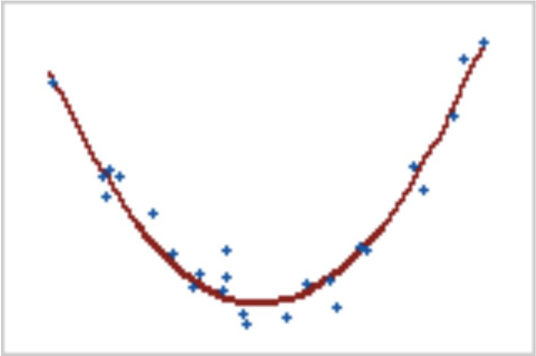
Pearson = −0.799, Spearman = −1

https://support.minitab.com/en-us/minitab/21/help-and-how-to/statistics/basic-statistics/supporting-topics/correlation-and-covariance/a-comparison-of-the-pearson-and-spearman-correlation-methods/

How exactly does Spearman differ from Pearson correlation?



Pearson = −0.093, Spearman = −0.093



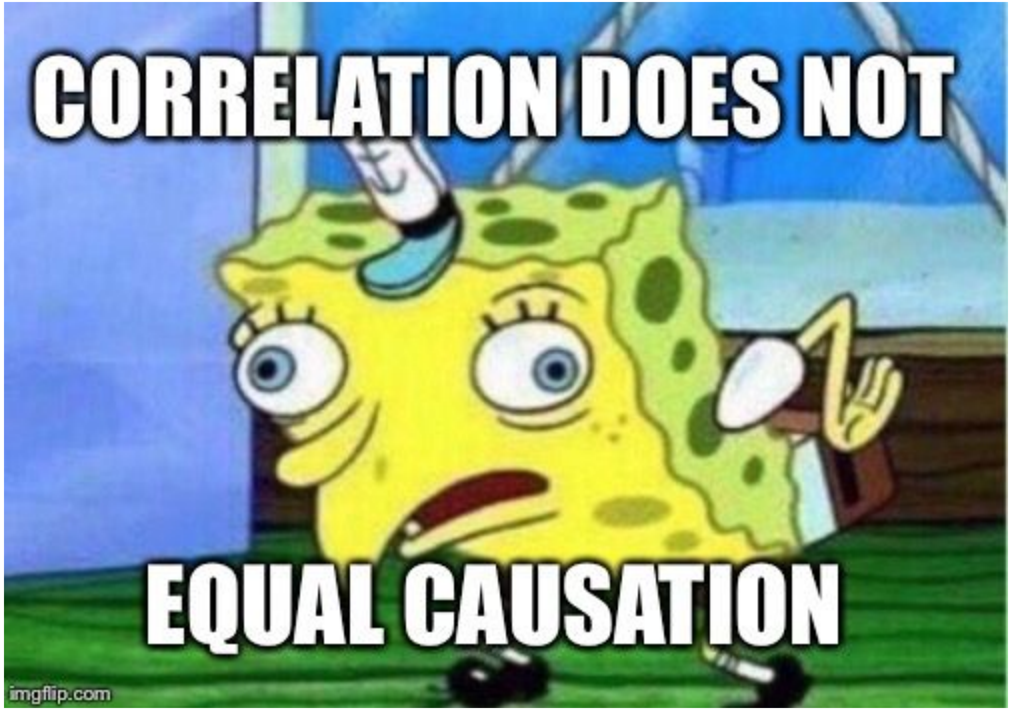
Pearson = 0, Spearman = 0

https://support.minitab.com/en-us/minitab/21/help-and-how-to/statistics/basic-statistics/supporting-topics/correlation-and-covariance/a-comparison-of-the-pearson-and-spearman-correlation-methods/

Takeaways on rank transforms

- Ranking your data can be useful, depending on the application
- If you want to compare rankings, use Spearman correlation to understand the monotonicity relationship between variables
- Monotonicity != Linear relationships
- And…

Still true, regardless of corr metric



https://www.lancaster.ac.uk/stor-i-student-sites/ziyang-yang/2021/02/26/statistics-in-social-science-1-how-to-choose-an-appropriate-statistical-test/

Admin

- Friday discussion: going over the prelim solutions, introducing Phase 2
- Phase 2 due Oct 19th

Admin

- Friday discussion: going over the prelim solutions, introducing Phase 2
- Phase 2 due Oct 19th
- HW4 posted Oct 11th, due Oct 26th
- **No homework over fall break! Enjoy :)**