Statistics Bootcamp Day 4

19 September 2019

Welcome to bootcamp!

Our goals:

1. Increase students' understanding of and confidence with basic statistical concepts.
2. Build students' programming intuition and data management skills.
3. Encourage collaboration and camaraderie among the graduate student cohort.

Overview of the week

~~Monday: mindset, descriptive & inferential statistics, summary statistics, and Stata workshop~~

~~Tuesday: graphing, exponents/logarithms, sampling distributions, and statistical significance~~

~~Wednesday: probability basics, file structure and data workflow~~

Thursday: variable types, functions, lines of best fit, prediction equations

Friday: matrix algebra basics, reading calculus

Today’s learning objectives

...categorize variables according to the type of variable.

...understand the difference between a function and a relation.

...explain the meaning of the slope of a line.

...draw a line of best fit and justify its location.

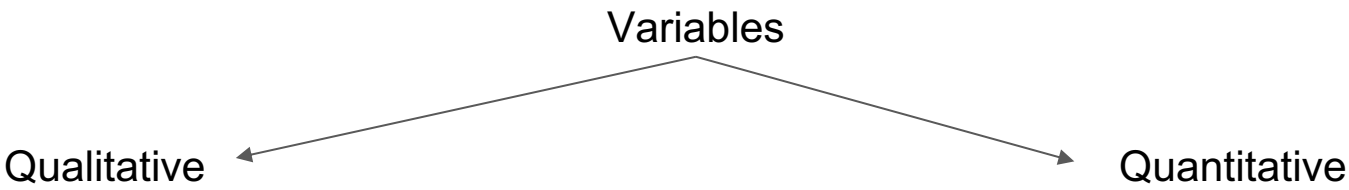
...describe the relationship between two variables in words, graph form, and equation form.

...understand the purpose of prediction equations and how to use them.

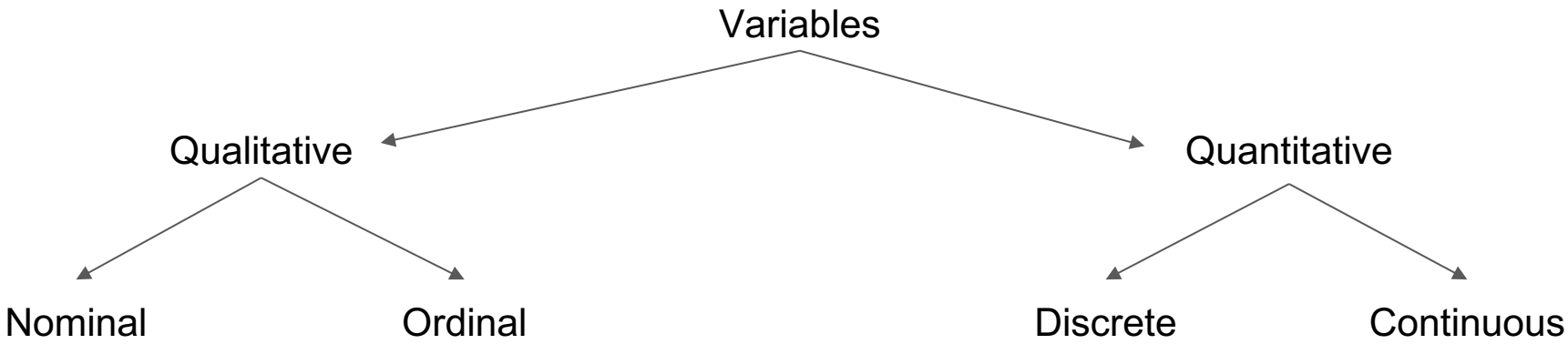
A reminder about variables

- A **variable** is something we measure for every **unit** in our sample.
- Variables are classified according to their type.

Types of variable



Types of variable



Types of variables (NOIR)

Qualitative	**Nominal**	Categories with no distinguishable order E.g. hair color, U.S. state, race

Types of variables (NOIR)

Qualitative	**Nominal**	Categories with no distinguishable order E.g. hair color, U.S. state, race
	Ordinal	Ordered categories E.g. Frequency measures, Likert scales

Types of variables (NOIR)

Qualitative	**Nominal**	Categories with no distinguishable order E.g. hair color, U.S. state, race
	Ordinal	Ordered categories E.g. Frequency measures, Likert scales
Quantitative	**Interval**	Numbers without a meaningful zero. The “interval” or distance between two numbers makes sense, but fractions do not E.g. temperature (in F or C), SAT scores

Types of variables (NOIR)

Qualitative	**Nominal**	Categories with no distinguishable order E.g. hair color, U.S. state, race
	Ordinal	Ordered categories E.g. Frequency measures, Likert scales
Quantitative	**Interval**	Numbers without a meaningful zero. The “interval” or distance between two numbers makes sense, but fractions do not E.g. temperature (in F or C), SAT scores
	Ratio	Continuous numbers with a meaningful zero. The interval between two numbers makes sense, as do fractions. E.g. income in dollars, height in inches

Note: these are sometimes called “**levels of measurement**”. They build on each other: all ordinal variables also fit the criteria for nominal, and so on.

Let’s categorize some variables!

Variable	**Quantitative/ Qualitative?**	**Discrete/ Continuous?**	**NOIR**	**How might Stata store this information (e.g. numeric, numeric with labels, or string)?**
Number of email messages				
Cost of textbooks (in dollars)				
Monthly cellular bill (in dollars)				
Level of happiness				

Variable	**Quantitative/ Qualitative?**	**Discrete/ Continuous?**	**NOIR**	**How might Stata store this information (e.g. numeric, numeric with labels, or string)?**
Number of email messages	Quantitative	Discrete	Ratio	Numeric
Cost of textbooks (in dollars)	Quantitative	Continuous	Ratio	Numeric
Monthly cellular bill (in dollars)	Quantitative	Continuous	Ratio	Numeric
Level of happiness	Qualitative or quantitative	Discrete	Ordinal or interval	Numeric (with or without labels)

Variable	**Quantitative/ Qualitative?**	**Discrete/ Continuous?**	**NOIR**	**How might Stata store this information (e.g. numeric, numeric with labels, or string)?**
Opinion about a woman's right to abortion (Likert scale of 1 to 7)				
Plan to vote (yes or no)				
Number of days per week that you use the library				
Calendar year				

Variable	**Quantitative/ Qualitative?**	**Discrete/ Continuous?**	**NOIR**	**How might Stata store this information (e.g. numeric, numeric with labels, or string)?**
Opinion about a woman's right to abortion (Likert scale of 1 to 7)	Qualitative or quantitative	Discrete	Ordinal or Interval	Numeric (with or without labels)
Plan to vote (yes or no)	Qualitative	Discrete	Nominal	Numeric (dummy variable, with or without labels)
Number of days per week that you use the library	Quantitative	Discrete	Ratio	Numeric
Calendar year	Quantitative	Discrete	Interval	Numeric

Why do we care about the type of variable?

The level of measurement of our variable tells us how we can analyze the variable (and also how the variable should be coded/stored in the data).

Today we're going to talk about scatter plots and lines of best fit: these work best for continuous, quantitative data.

Relations vs. Functions

How can we think about the connection between two different variables?

We can categorize the connection as either a **relation** or a **function**. For example, let's think of one variable as the **input** and the other as the **output**.

If there's any output for a given input, it's a relation.

If there's only one output for every input (e.g., we can solve for y =) the connection is a FUNCTION. (all functions are relations, but not all relations are functions!)

If you have a function and are given an input, you KNOW what the output will be.

Relations vs. Functions: Examples

x	y
0	0
1	1
2	2
3	3
4	4

This is a function (and a relation).
y = x

x	y
0	0
1	1
2	2
2	3
4	4

This is a relation.

x	y
0	3
1	3
2	3
3	3
4	3

This is a function (and a relation).
y = 3

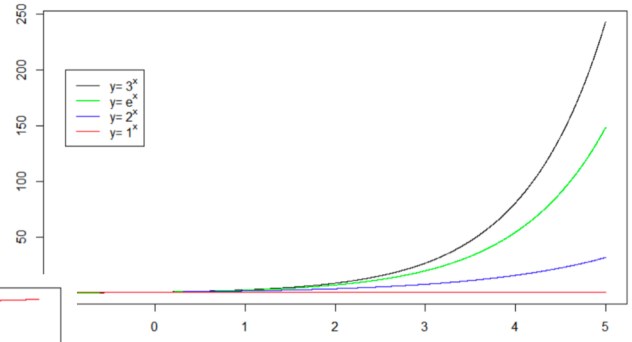
Some common functions...

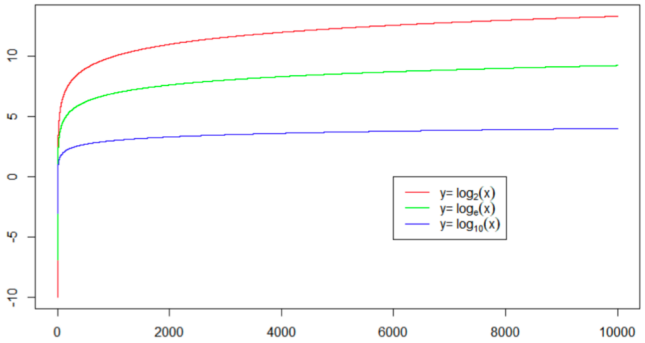
Exponents a^x

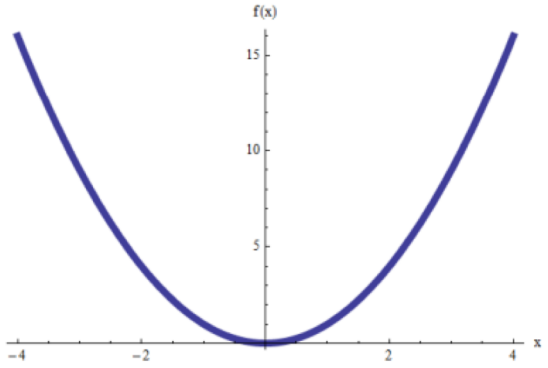
Logarithms log(x)

Polynomials

$y = ax + bx^2$







Why do we care about functions?

With inferential statistics, we use functions between two (or more) variables to predict the values of an **outcome variable** (or something we want to predict).

Different functions have different SHAPES. Sometimes our data follow different patterns, so we use different functions/patterns to model them (e.g. a squared term x^2 to have a curved prediction line).

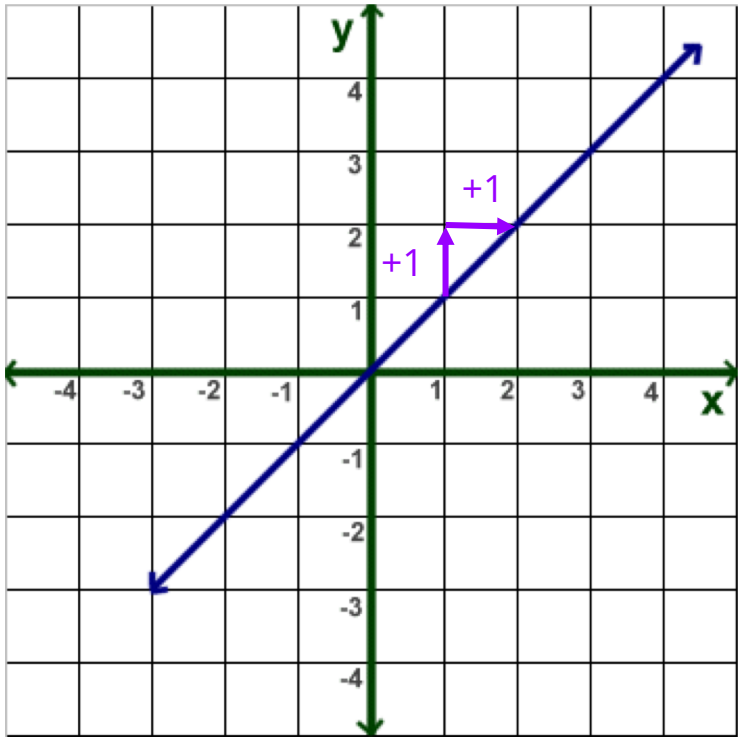
This process involves creating **prediction equations** that mix together different variables and functions. We'll return to this later and in 381!

Activity

In pairs...

- Identify the pattern
- Describe the pattern using a **function**
- Predict how many blocks will be in the 30th figure

Equation of a Line



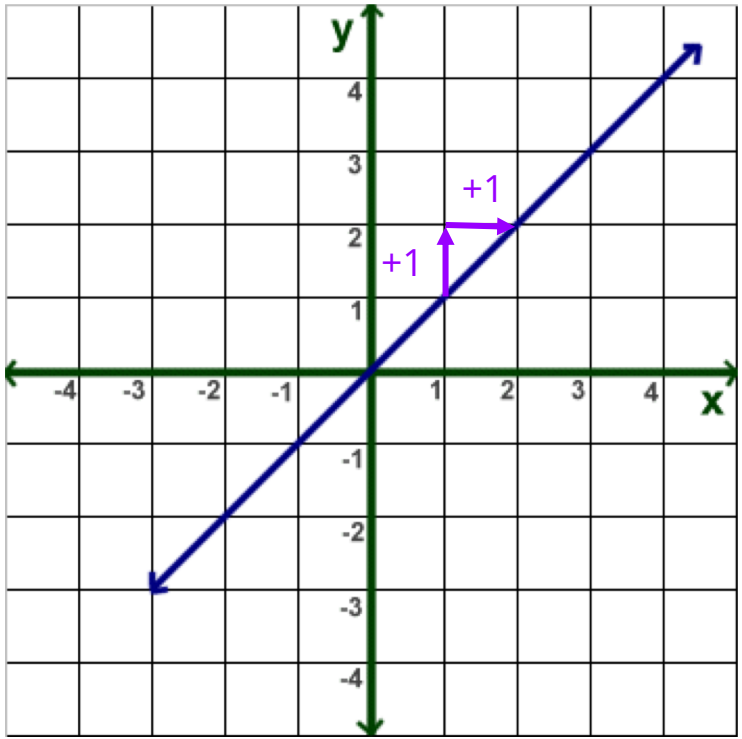
m = 1 / 1
Y = 1X +0
Y = X

slope

$$Y = mX + b$$

y-intercept

Equation of a Line



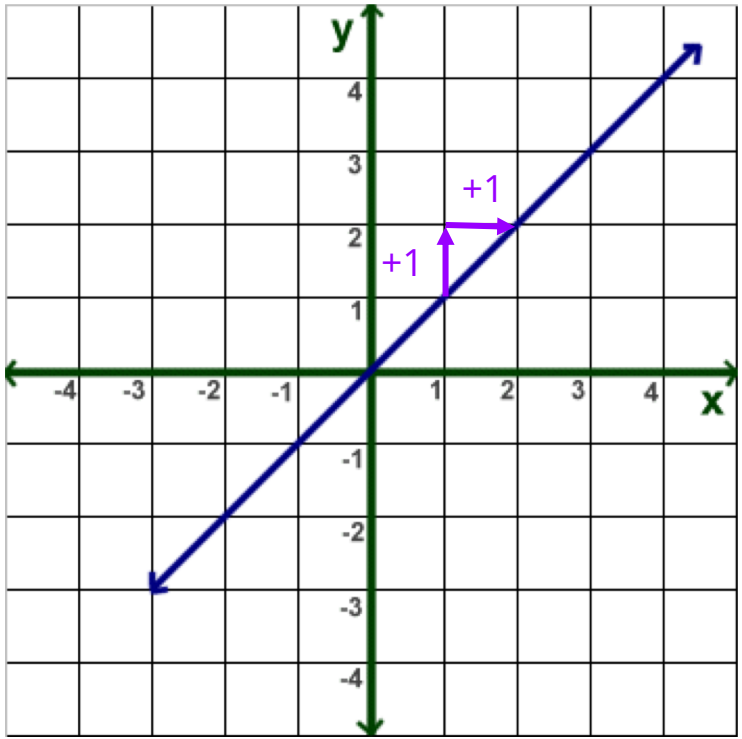
m = 1 / 1
Y = 1X +0
Y = X

slope = rise / run = change in Y / change in X

$$Y = mX + b$$

y-intercept = value of Y when X is 0 = where the line crosses the Y axis

Equation of a Line



m = 1 / 1
Y = 1X +0
Y = X

For every 1 unit increase in X,
there is a 1 unit increase in Y

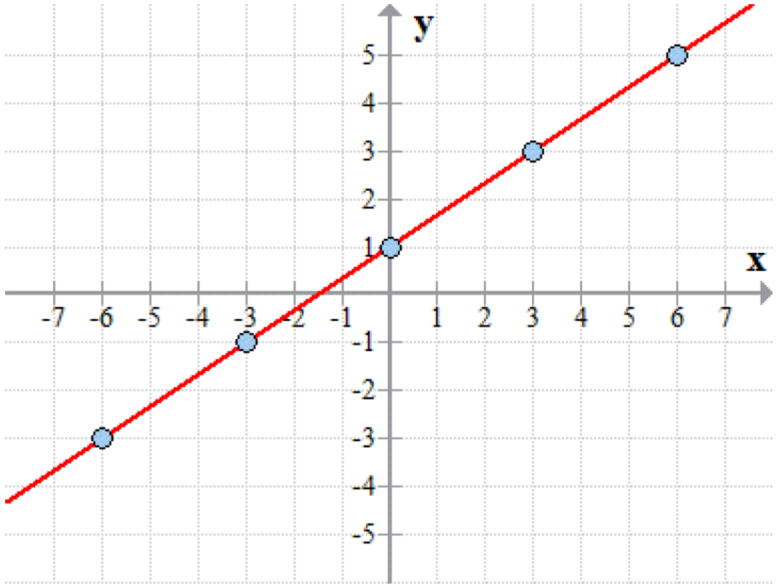
slope = rise / run = change in Y / change in X

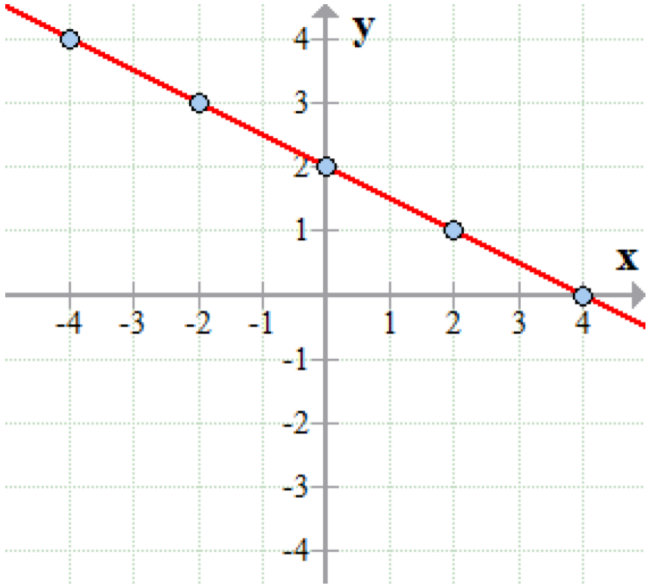
$$Y = mX + b$$

y-intercept = value of Y when X is 0 = where the line crosses the Y axis

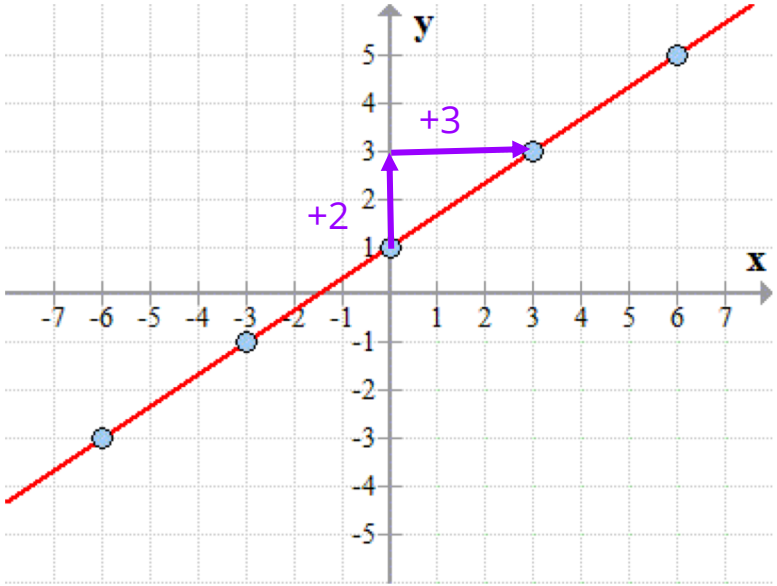
When X is 0, Y is 0

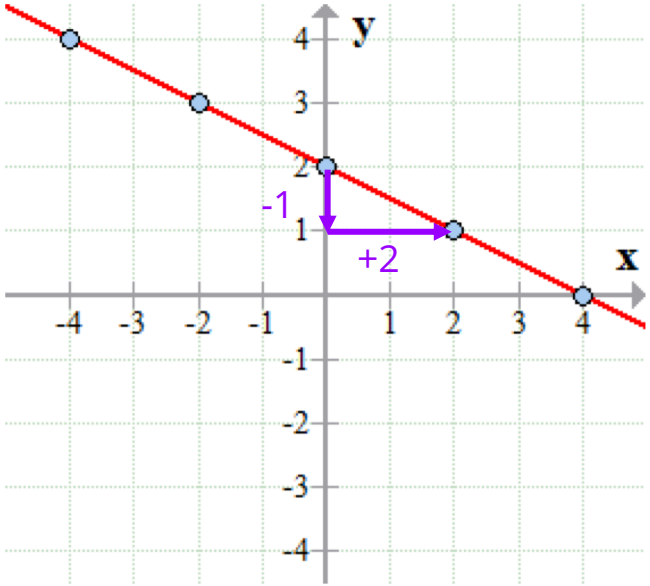
Equation of a Line



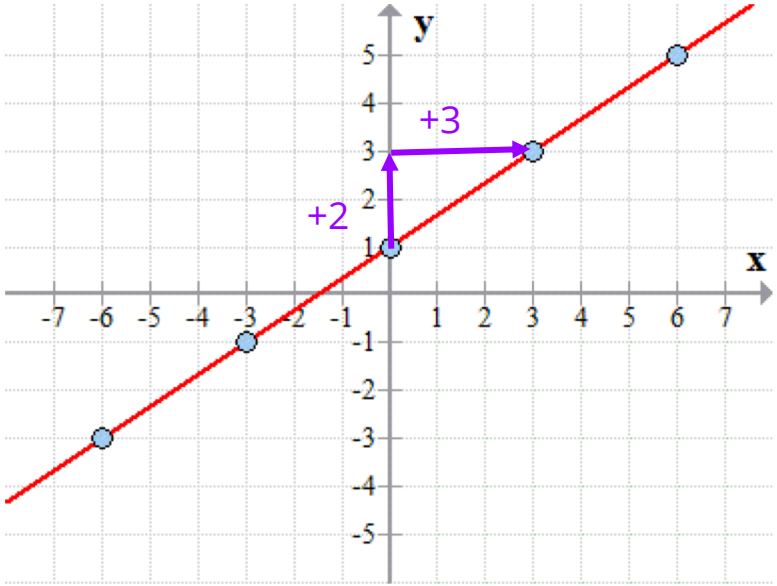


Equation of a Line

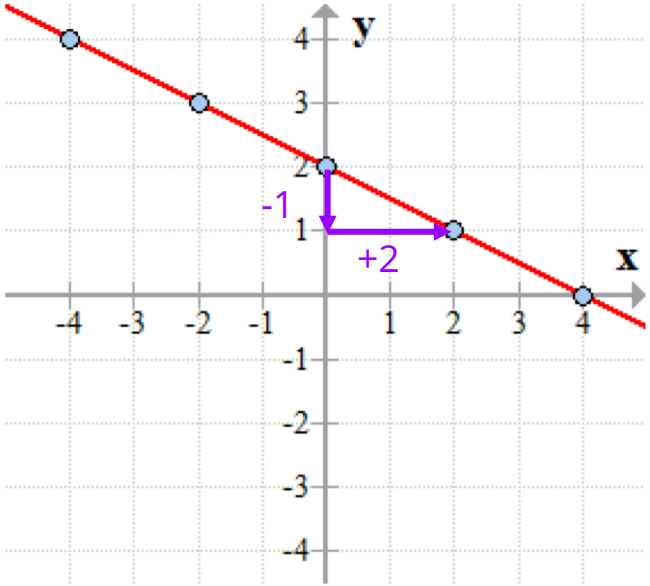




Equation of a Line

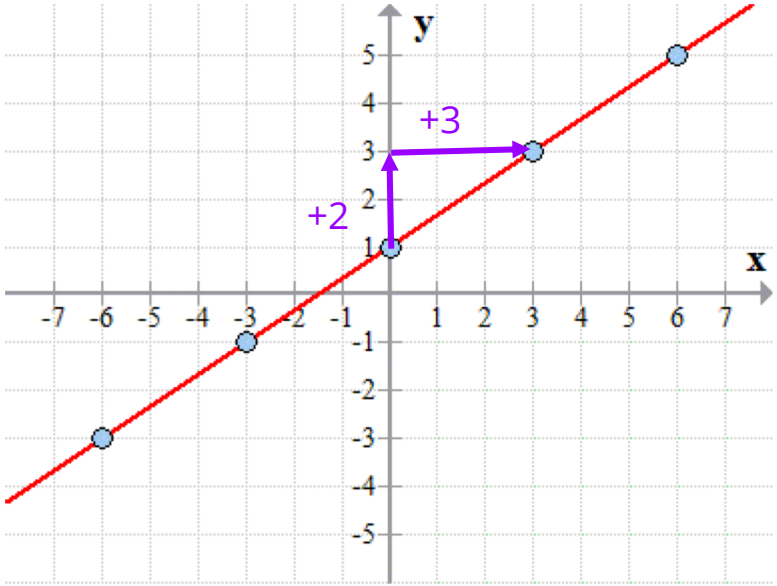


m = 2 / 3
Y = 2/3X +1



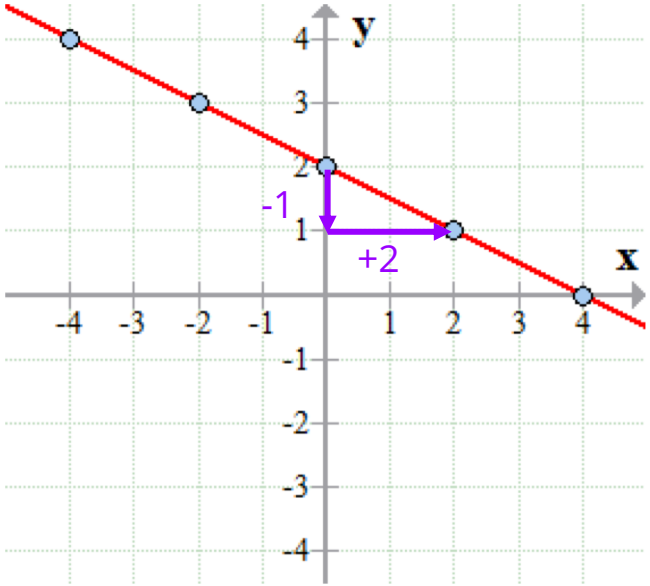
m = -1 / 2
Y = -1/2X +2

Equation of a Line



For every 1 unit increase in X, there is a 2/3 unit increase in Y

m = 2 / 3
Y = 2/3X +1



m = -1 / 2
Y = -1/2X +2

For every 1 unit increase in X, there is a 1/2 unit decrease in Y

Activity

Use the equation for each pattern and draw the line on the board.

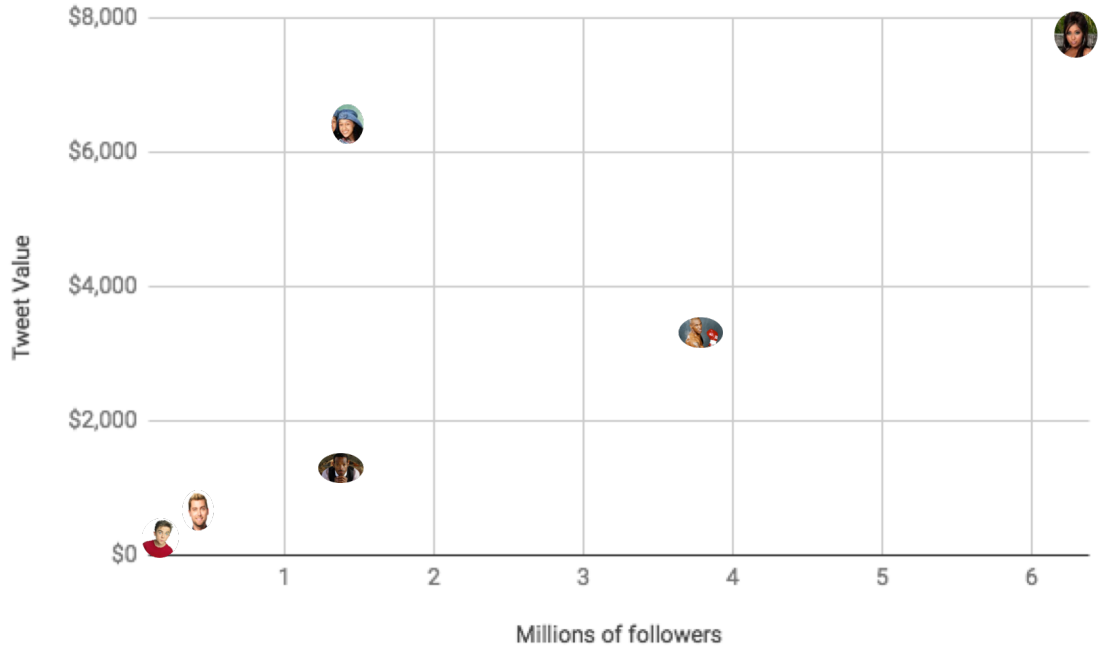
Interpret the slope and y-intercept.

lunch

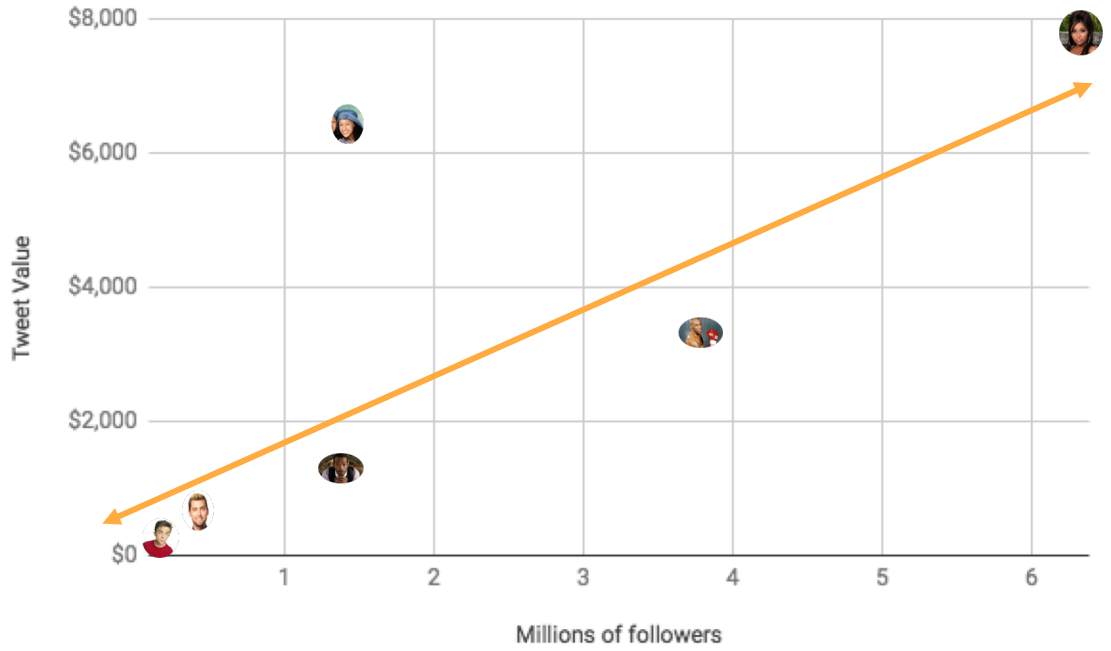
B-list celebrities and sponsored tweets

Celebrity	Twitter Followers (Millions)	Tweet Value
Lance Bass 	0.45	$650
Tia Mowry	1.45	$6,500
Snooki 	6.3	$7,800
Mike Tyson	3.8	$3,250
Marlon Wayans 	1.36	$1,300
Frankie Muniz 	0.18	$252

Where’s the line of best fit?

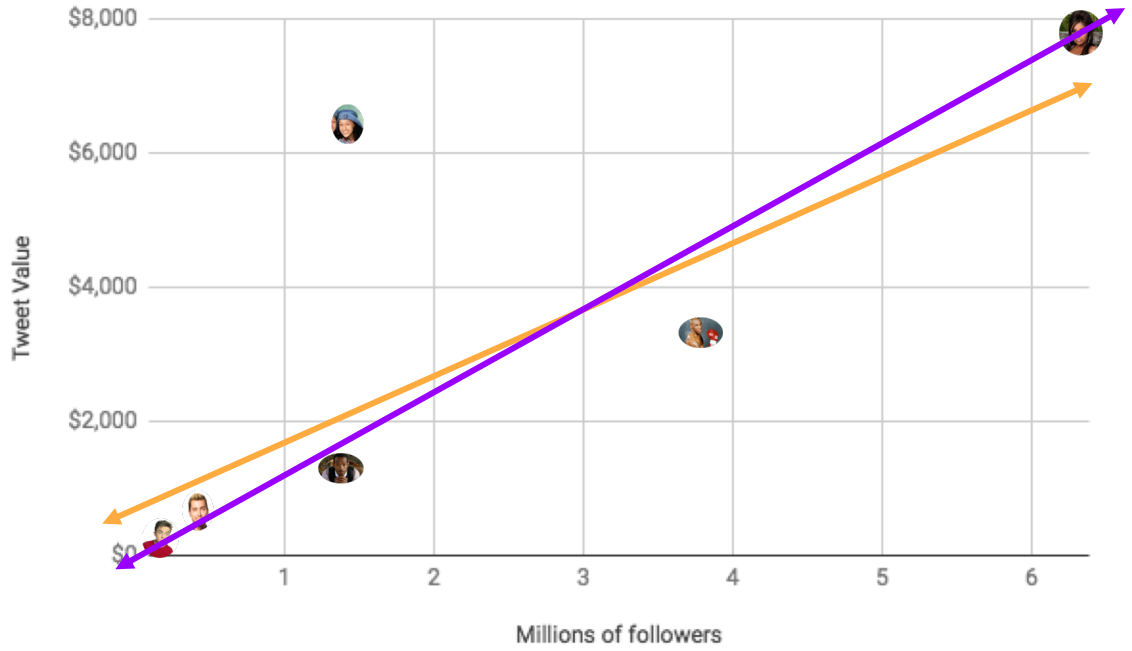


Where’s the line of best fit?



Y = 1025 X + 980
Dollars per tweet = 1025 (Millions of followers) + 980

Where’s the line of best fit?

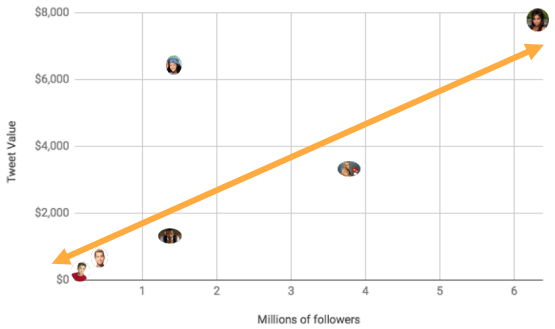


Y = 1025 X + 980
Dollars per tweet = 1025 (Millions of followers) + 980

Using the line of best fit

If Melissa Joan Hart has 720,000 followers, how much money per tweet can would you predict her to make?

What about Katy Perry, who has 103 million followers?



Y = 1025 X + 980

Dollars per tweet = 1025 (Millions of followers) + 980

Equation modeling in statistics/sociology

In sociology, we often care about the relationship between two things/variables.

We model this relationship using a function/**equation**/line of best fit!

Sometimes we are just looking at the relationship between two variables, but oftentimes we incorporate lots of different variables. In other words, we use many different **predictors** (independent variables) to predict our **outcome** (dependent variable).

This is called **regression analysis**!

Ordinary Least Squares (OLS) Regression

OLS regression uses a *prediction equation* to predict values of an outcome variable.

Here are a bunch of ways to write prediction equations (you'll see these soon!):

$$Y = a + bX$$

$$Y = \beta_0 + \beta_1 X$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + e_i$$

$$\hat{y}_i = \beta_0 + \beta_1 X_1$$

A closer look…

The regression coefficient(s)

The dependent variable, or the outcome

$$Y = \beta_0 + \beta_1 X$$

The independent variable(s), or the predictor(s)

The intercept/constant

A closer look…

The regression coefficient(s)

The dependent variable, or the outcome

$$Y = \beta_0 + \beta_1 X$$

The independent variable(s), or the predictor(s)

The intercept/constant

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + e_i$$

The error term!! Our equation never predicts the outcome perfectly. There's always random error.

This is **multiple regression**

Note that we can have multiple predictors, each with its own regression coefficient

A closer look…

The regression coefficient(s)

The dependent variable, or the outcome

$$Y = \beta_0 + \beta_1 X$$

The independent variable(s), or the predictor(s)

The intercept/constant

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + e_i$$

The error term!! Our equation never predicts the outcome perfectly. There's always random error.

Note that we can have multiple predictors, each with its own regression coefficient

Called "y-hat". Predicted value of the outcome.

$$\hat{y}_i = \beta_0 + \beta_1 X_1$$

(no error term means that the equation is showing the predicted value of Y, not the true value.)

Regression in Stata with one predictor

```
. regress yrsed female if age >=25 & age <=34
```

Source	SS	df	MS		
Model	276.742433	1	276.742433	Prob > F	0.0000
Residual	156979.922	18,536	8.46892111	R-squared	0.0018
Total	157256.664	18,537	8.48339343	Root MSE	2.9101

Number of obs	=	18,538
F(1, 18536)	=	32.68
Prob > F	=	0.0000
R-squared	=	0.0018
Adj R-squared	=	0.0017
Root MSE	=	2.9101

yrsed	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	.2444469	.0427623	5.72	0.000	.1606289	.3282649
_cons	13.31212	.0306297	434.62	0.000	13.25208	13.37216

One predictor - How would we write the equation?

```
. regress yrsed female if age >=25 & age <=34
```

Source	SS	df	MS		
Model	276.742433	1	276.742433	Number of obs	= 18,538
Residual	156979.922	18,536	8.46892111	F(1, 18536)	= 32.68
				Prob > F	= 0.0000
				R-squared	= 0.0018
				Adj R-squared	= 0.0017
Total	157256.664	18,537	8.48339343	Root MSE	= 2.9101

yrsed	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	.2444469	.0427623	5.72	0.000	.1606289	.3282649
_cons	13.31212	.0306297	434.62	0.000	13.25208	13.37216

$$Y = \beta_0 + \beta_1 X$$

One predictor - How would we write the equation?

```
. regress yrsed female if age >=25 & age <=34
```

Source	SS	df	MS	Number of obs	=	18,538
				F(1, 18536)	=	32.68
Model	276.742433	1	276.742433	Prob > F	=	0.0000
Residual	156979.922	18,536	8.46892111	R-squared	=	0.0018
				Adj R-squared	=	0.0017
Total	157256.664	18,537	8.48339343	Root MSE	=	2.9101

yrsed	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	.2444469	.0427623	5.72	0.000	.1606289	.3282649
_cons	13.31212	.0306297	434.62	0.000	13.25208	13.37216

$$Y = \beta_0 + \beta_1 X$$

$$\widehat{yrsed} = 13.31 + 0.24 * female$$

OLS Regression Income on Age and Education

	Wage and salary income		
	(1)	(2)	(3)
Age	-7.044	3,252.0***	2,891.9***
	(11.65)	(95.60)	(90.64)
Age Squared		-37.34***	-32.46***
		(1.087)	(1.031)
Years of Education			3561.5***
			(39.98)
Constant	27,764.5***	-39,131.1***	81,235.3***
	(511.3)	(2,012.7)	(1,964.3)
R-squared	0.000	0.017	0.118

Standard errors in parentheses
* p<0.05 ** p<0.01 *** p<0.001
Source: CPS 2000
N = 69,305

Example from CPS data

REGRESSION IN THE REAL WORLD

Table 1: Mixed Models Predicting Mental Health and Treatment

	(1) Self-reported mental health	(2) Treatment
Variance Components of Random Effects		
Cohort	0.013	0.009
Period	0.030	0.013
Individual	12.60	
Individual-level Characteristics (Fixed Effects)		
Age	−.020***	−.019***
	(.001)	(.001)
Age-squared	−0.000***	−.001***
	(0.000)	(0.000)
Female	.464***	.213***
	(.009)	(.011)
White	.394***	.508***
	(.012)	(.014)
Hispanic	−.285***	−.438***
	(.013)	(.017)
General health (standardized)	−1.209***	−.251***
	(.005)	(.006)
Married	−.502***	−.567***
	(.010)	(.012)
College-educated	−.242***	.548***
	(.010)	(.012)
Below poverty line	.968***	.208***
	(.014)	(.015)
Treatment	3.485***	
	(.018)	
Has health insurance		.819***
		(.017)
K6 score		.156***
		(.001)
Constant	2.139***	−4.357***
	(.050)	(.045)
Observations	601,237	601,237

*P < .05 **P < .01 ***P < .001. Standard errors in parentheses.
Source: NHIS 1997-2017
Note: self-reported mental health is measured using the K6 index, which ranges from 0 to 24. Coefficients in Model 2 are in log odds. Age and age-squared are grand-mean centered.

Fancy example from my research

stata time

stata time

stata time

EXIT TICKET

1. On a scale of 1-5, how ready do you feel to start 381?

2. Any thoughts on today's lesson? Points of confusion/things you liked?

3. What's a burning question you have about grad school that you'd like us to answer?