2017 GSPIA Amazing Analytics Race

Wednesday Training Camp

Sera Linardi

Assistant Professor of Economics

9am Getting ready: Your To-Do List

Introductions

Tekky Bambang and Quintin Lemom (IT) and TAs Alicia Houser and Meghan Yost.

Introduce yourself to 2 **new** people around you.

1. Register and find your ID number on your name tag
2. Get STATA if you haven't already.
3. Get online if you haven't already. If you are unable to get online, request paper copies of exercises from the TAs and move closer to the front to see the slides.
4. Go to http://www.linardi.gspia.pitt.edu/?page_id=564. The SCHEDULE of the day is online for you to check at any time.
5. Create a folder in your computer for all your files for math camp. Download all materials for Wednesday into that folder
6. Open STATA, go to File, Change Working directory to your math camp folder.
7. Click on the Baseline math survey and try it. Use the ID # from your name tag.

We will start lecture as soon as everyone is set up (between 9:15-9:30 am).

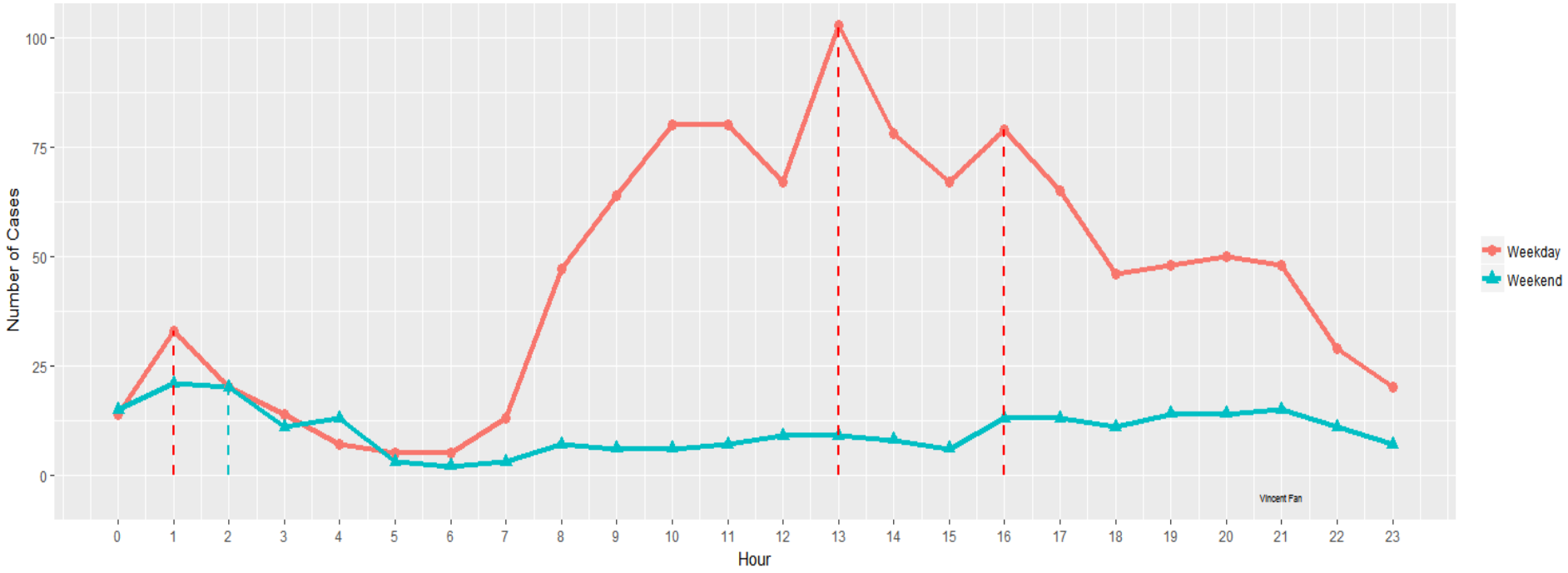
Welcome

- Your instructor:
- Sera Linardi (linardi@pitt.edu)
- PhD in Social Science, California Institute of Technology
- Behavioral economist; I use experiments to study the psychology behind the motivation to help others (e.g in the context of interethnic relationships in Afghanistan, lending in Islamic banking) and behind the use of social services.

- I teach Quant II in the Fall and Game theory/Behavioral Economics in the Spring. I'm also teaching data visualization with R in Spring.

R Data Visualization Capstone

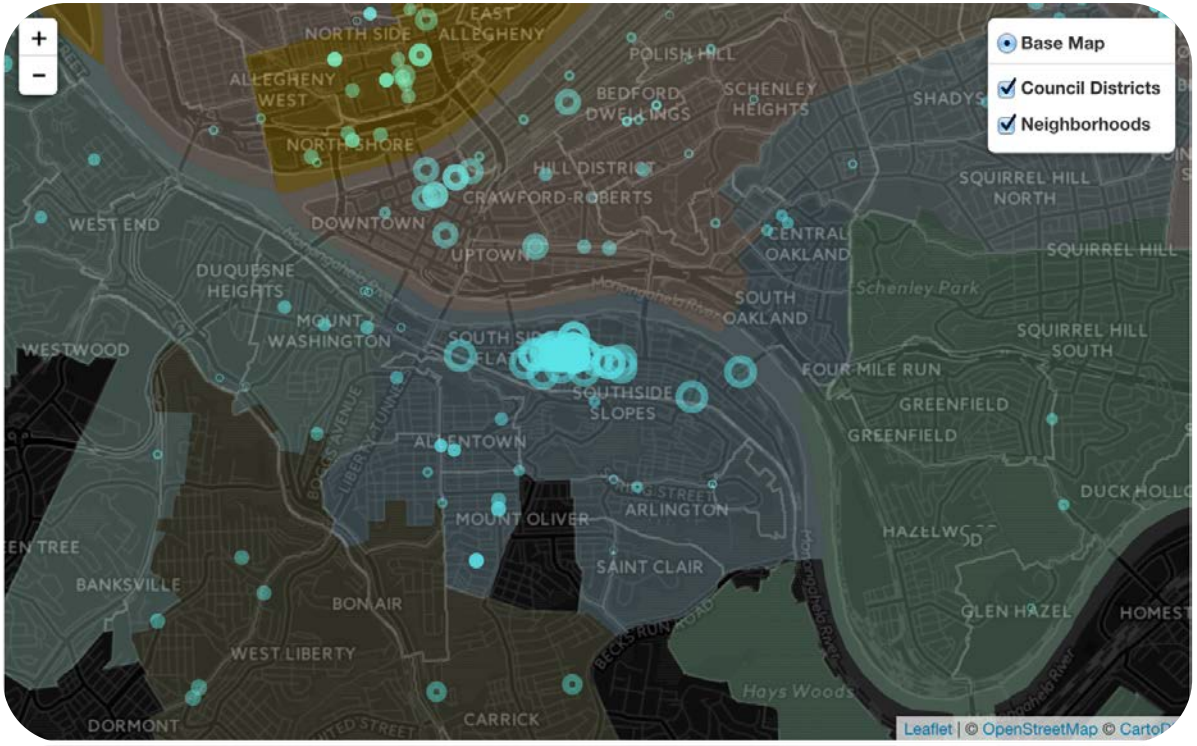
- **Daily Distribution of Juvenile Delinquency Among Weekdays and Weekends**



R Data Visualization Capstone

Our Map

GIS for Arrest Data https://visiuchen.shinyapps.io/PITTSBURGH-CRIME-MAP/



What this workshop is and is NOT

What are we doing today? We are beginning your GSPIA journey with the end in mind: a career solving real world problems

First, let's define what this workshop will NOT do:

- Guarantee you an A in Quant I or Micro or any quant class
- Make you a math whiz
- Explain any mathematical concept in depth

<u>What this workshop aims to do:</u>

- Connect quant methods to the real world.
- Give you a previous of ALL the math you will see during your time here. You will most likely not encounter any math that you have not seen today.
- Provide a quick-and-dirty, hands-on experience of how quant methods give you an additional edge in tackling policy questions

Schedule and people you will meet today

- 9:00 registration, materials & setup, baseline quiz.
- 9:15-10:20 Lecture 1: Linear functions, Exercise 1
- 10:20-10:30 Meet your quant professors
- 10:30-10:45 Break
- 10:45-12:00 Lecture 2: Nonlinear functions and derivatives, Exercise 2
- 12:00-1:00 Break
- 1:00-2:30pm Amazing Analytics Race teams (TAs), Lecture 3: Intro to Stats, Exercise 3
- 2:30-2:45pm Break
- 2:45pm-3:30pm Team exercise

And.. what is GSPIA's Amazing Analytics Race ?

- At the end of today, you will be randomly split into pairs for tomorrow.
- Your mission will be explained tomorrow: you will have 3 hours to solve a puzzle by interlocking a series of 10 clues with your partner.
- You will use real world data, the quantitative methods you learn today, and lots of creativity.
- What's at stake: 1st place team = a $200 Bookstore gift certificate. 2nd place team = $100. 3rd place = $50.

- After teams are formed today, we will brief you on the rules of the race, and your team will get to practice working together.

How today's training camp works

- Data - Lecture (<1hr)– Exercise (10 mins) – Review the exercise (5-10 mins)

- You have the slides on your computer, so you can always go back / make notes, etc.

- Ask questions! There is no dumb question, this **is** a refresher workshop so forgetting basic stuff is totally okay. In completing exercise feel free to ask your neighbors/TAs/instructor for help.

- Please don't browse the internet/ phone for unrelated stuff. If you are waiting for others to finish, see if anyone near you needs help, or try new things with STATA.

Imagine you are an advisor to the mayor of Pittsburgh

- He is wondering whether or not to approve 10 new businesses on a strip of a crowded highway: businesses bring jobs but worsen congestion

- What you have to help you advise him:
  - Data on travel time on several highways given the number of cars on the highway (Cars.csv)
  - Data on number of cars given number of businesses along the highway (Business.csv)
  - Public opinion expert's estimated relationship between business development, traffic congestion and support for city government

Breaking down the question into mathematical concepts

1. how long does it take to travel the highway? (random variable)
2. how does the # (*number*) of cars affect travel time? (correlation, linear regression, slope)
3. can adoption of a different traffic system reduce congestion? (simultaneous equations)
4. how does the # of businesses affect # of cars?(nonlinear equations)
5. what is the optimal # of business to have? (optimization)

1. random variable

How long does it take to travel through the highway?

Random variable

How long does it take to travel 20 miles on a city highway at 8am in the morning? Hands = 20 mins, 30 mins, 40 mins

- Different day, same highway, same hour in day = different travel time.
- Statistics is learning to get the information out of this uncertainty.
- 'Time needed to travel' is a random variable = the value is subject to variation due to chance.

- Is what is written on this board ALL the possible travel times for 20 miles? No. That would be the *population*. This is a *sample*. We usually only observe a sample of realizations of the random variable of interest.

Distribution: what the population looks like

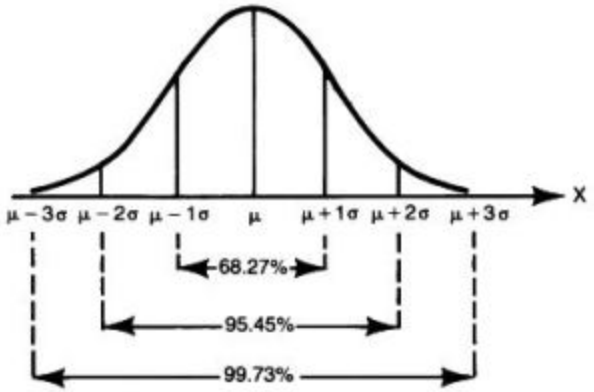
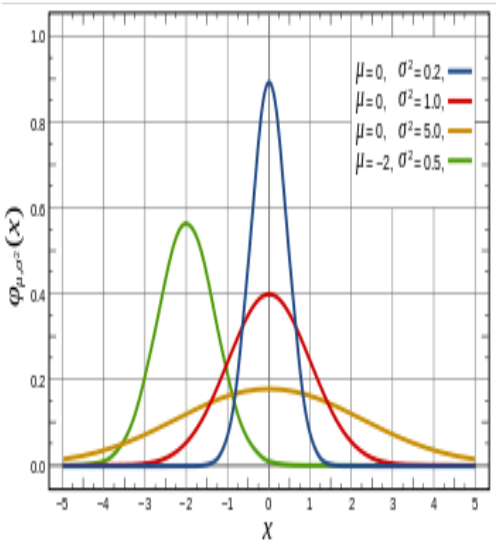


Figure 2



- Suppose this is all possible values of travel time and how likely you are to get any of them. Suppose the mean is 20 minutes, and the "standard deviation" is 5 minutes. Bigger standard deviation = more variability.
- This is a normal distribution: notice its symmetric about its mean.
- Things follow some simple rules with the normal distribution:
- Prob of being late when you give yourself 20 minutes is (100%) / 2 = 50%
- Prob of being late when you give yourself 20+5 minutes is (100%-68%) / 2 = 16%
- Prob of being late when you give yourself 20-5 minutes is 68%+16% = 84%
- Prob of being late when you give yourself 20+2*5 minutes is (100%-95.45%) / 2 = 2.275%

- Statistics is a study of random variable, its distribution, its relationship with other random variables, and what we can infer about populations from sample.

Looking at data

- Suppose cars.csv contains a random sample of travel time and # of cars on Pittsburgh highways.
- How do you load cars.csv into STATA so you can look at it?

- Loading with Data Editor. Open cars.csv in Excel. Highlight, copy. Open data editor. Click on first cell and paste. Treat first row as variable name.

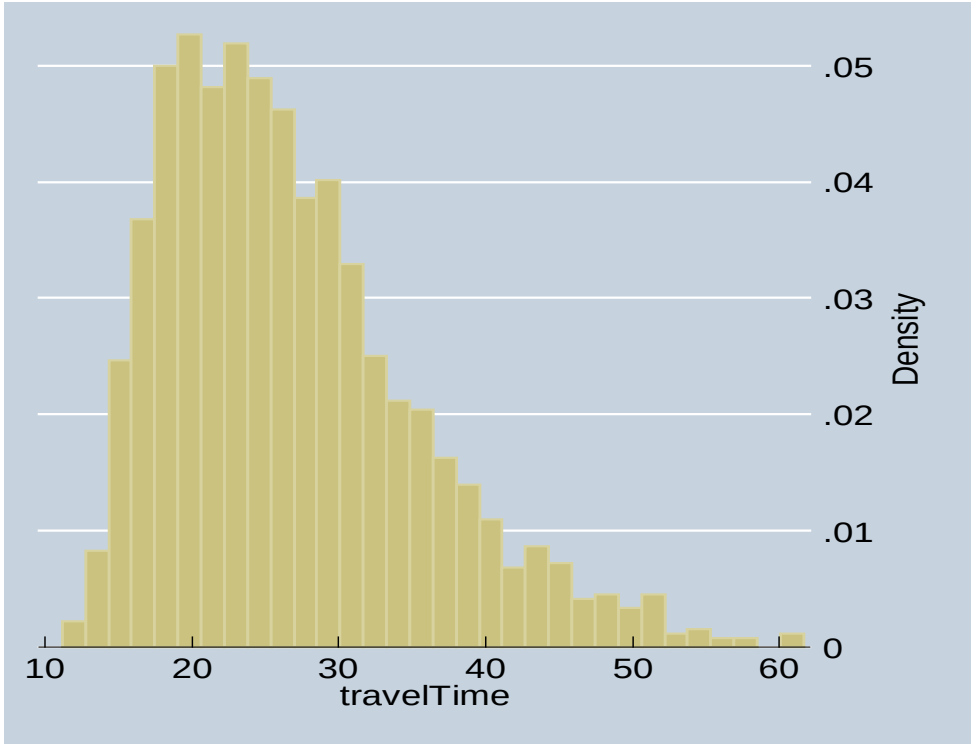
- (In general we'll use STATA in two ways today, first using the drop down menu, and then using code.)

- Note that the STATA and statistics you will learn today is just quick and dirty. You will learn how to use it properly in Quant I (with Jeremy) and Quant II (with me).

Travel time

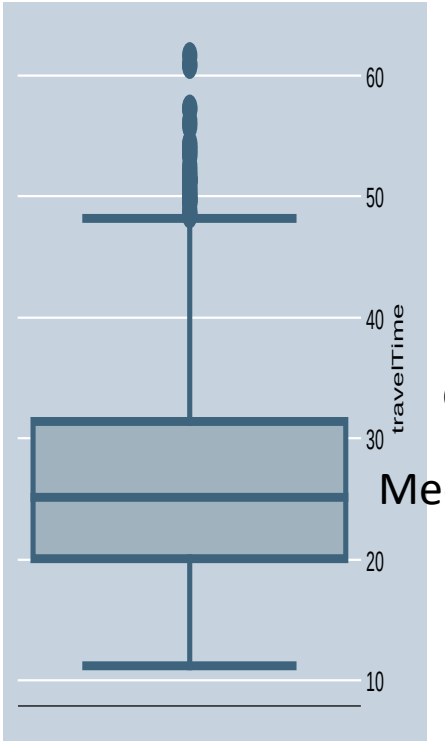
Mode, median, mean?

Histogram



hist traveltime (not normal, but we'll treat it as such today)

Boxplot



Upper quartile

Median

Lower quartile

graph box traveltime

Graphics → Histogram ->Variable: traveltime

```
. mean traveltime

Mean estimation                     Number of obs    =      1674
```

	Mean	Std. Err.	[95% Conf.	Interval]
traveltime	26.71808	.2103392	26.30553	27.13064

According to this sample, the average time needed to travel 20 miles on a
Pittsburgh highway is 26.7 minutes.

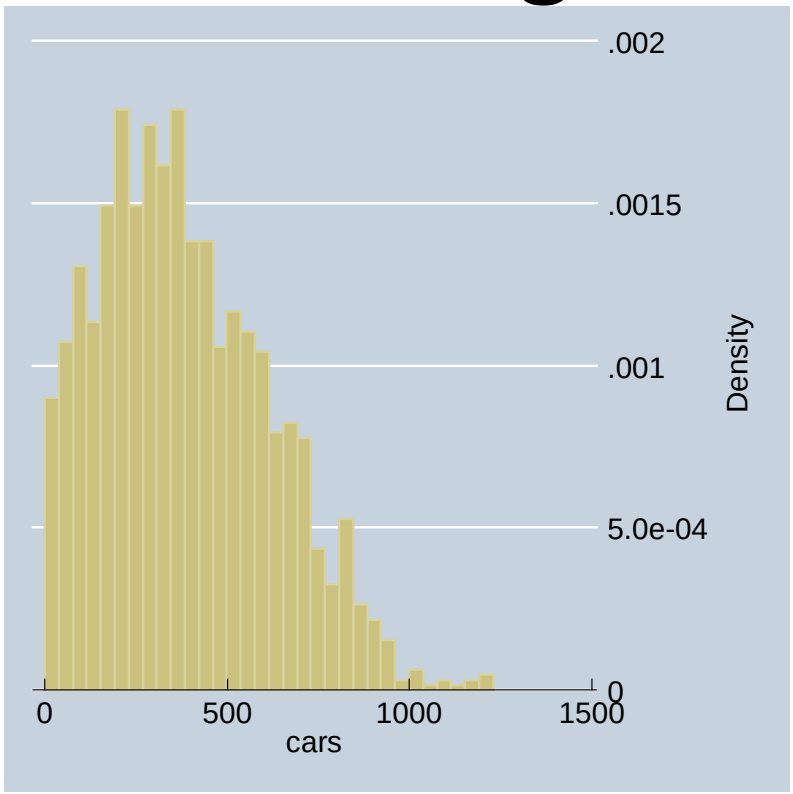
How confident can we be in this estimate?
In other words – if we take a different sample, will it also give the same average travel time?

The standard error tell us how much variation in travel time there is in the data – the smaller, the less variability.
The confidence interval combines the mean with the standard error; it tells us that we can be 95% confident that the true average time needed to travel 20 miles on a
Pittsburgh highway is between 26.31 and 27.13 minutes.

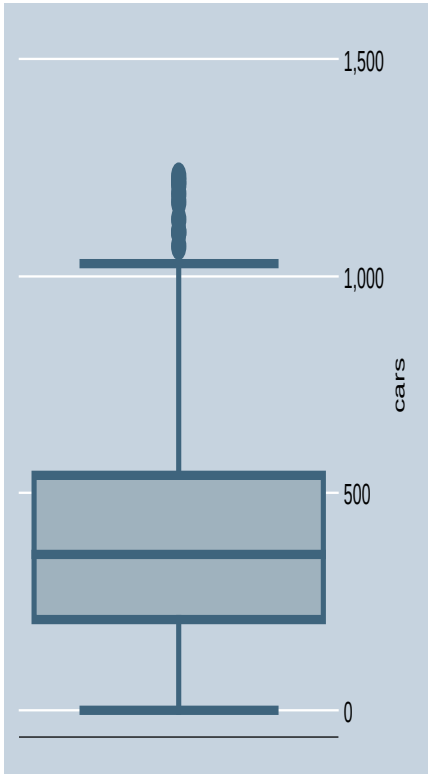
Given you're interested in congestion, you also look at the # of cars on the highway.

of cars on the highway

Histogram



Boxplot



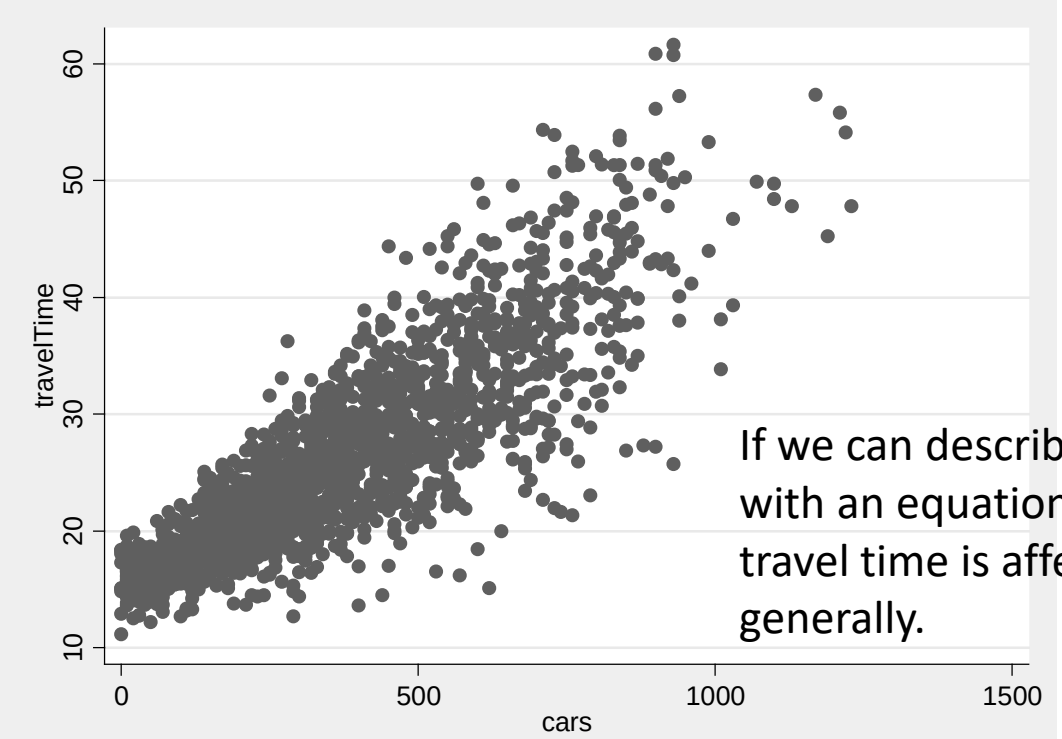
Hmm.. does this help you understand traffic congestion?

2. correlation, linear regression, slope / rate / derivative

how does the # of cars affect travel time?

Relationship between two random variables

correlation between travel time and # of cars

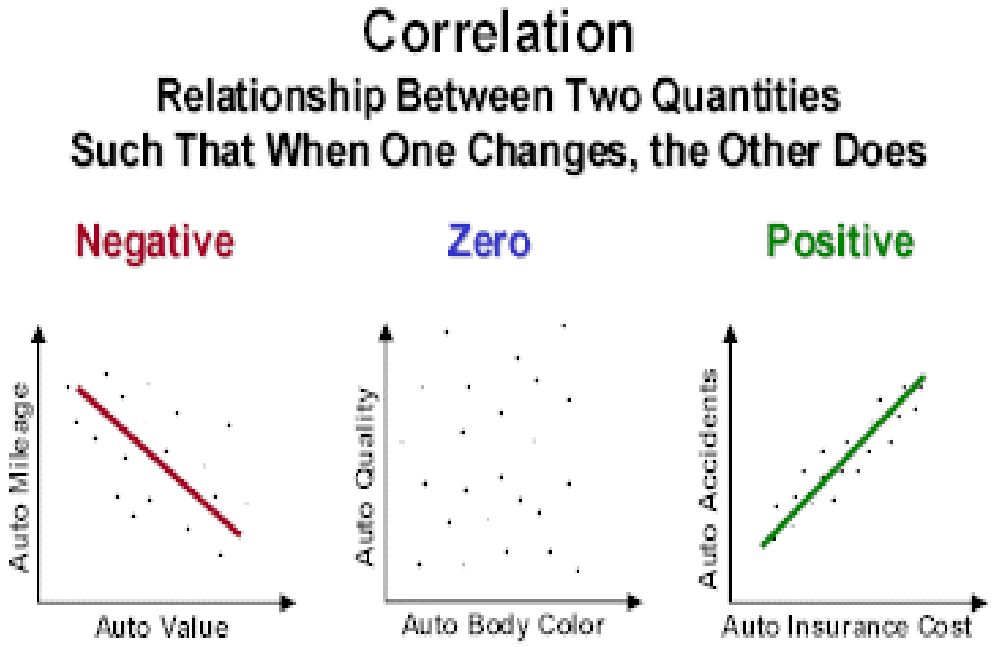


If we can describe this relationship with an equation, we can tell how travel time is affected by cars more generally.

scatter traveltime cars

Graphics → Twoway ->Create->Y variable: traveltime, X variable: cars

Scatterplot shows correlation between two variables.



To find the relationship, we can try to fit a line across this scatterplot that is the closest possible to ALL the points. This is a regression line.

Regression

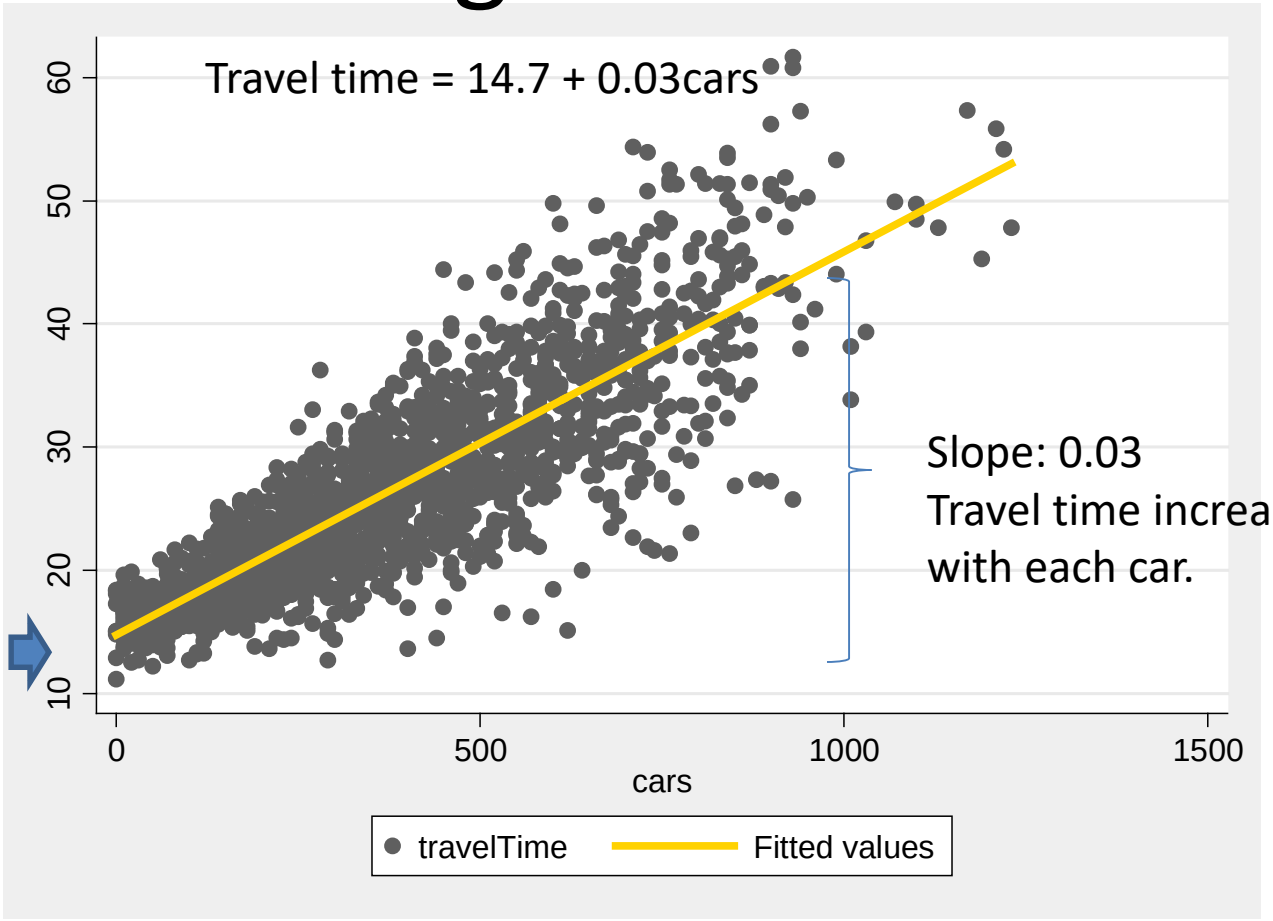
reg traveltime cars

```
------------------------------------------------------------------------------
  traveltime |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cars |   .0311483   .0004892    63.67   0.000     .0301888    .0321078
       _cons |    14.7139   .2201573    66.83   0.000     14.28208    15.14571
------------------------------------------------------------------------------
```

Statistics→ Linear model ->Linear Regression ->Dependent variable: traveltime, Independent variable: cars

traveltime = 14.7 + 0.03cars What does it mean?

Drawing a linear function



With an increase of 1000 cars, travel time increases by 1000*0.03 = 30 minutes. So with a thousand cars on the highway, total travel time is 14.7+30 = 44.7 minutes

With linear functions, an increase in X always increases Y by the same amount. For example, one additional car increase travel time by 0.03 minutes, regardless of whether there's 100 or 1000 cars on the freeway.

Returning to the regression

reg traveltime cars

```
------------------------------------------------------------------------------
  traveltime |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cars |   .0311483   .0004892    63.67   0.000     .0301888    .0321078
       _cons |    14.7139   .2201573    66.83   0.000     14.28208    15.14571
------------------------------------------------------------------------------
```

How confident can we be that travel time increases by 0.03 with each car?
The confidence interval tells us that we can be 95% confident that every car increases travel time by between 0.03 or 0.032 minutes.

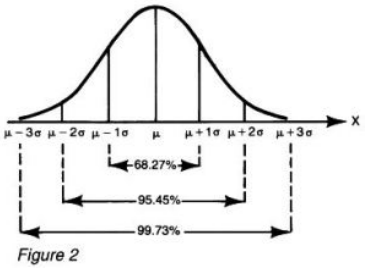


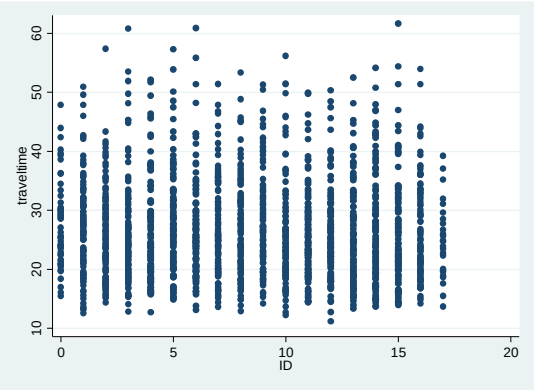
Figure 2

The p-value tells us that the probability of finding a coefficient of 0.03 in this data when there is actually no relationship between travel time and number of cars is 0.000.

In contrast, here is the relationship between the ID number of the person who is recording the data, and travel time.

reg traveltime v1

```
------------------------------------------------------------------------------
  traveltime |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ID |  -.0523915   .0433933    -1.21   0.227    -.1375025    .0327
       _cons |   27.15703   .4200033    64.66   0.000     26.33324    27.9
------------------------------------------------------------------------------
```



The p-value tells us that the probability of finding a coefficient of 0.052 in this data when there is actually no relationship between travel time and number of cars is 0.227.

By convention, the cutoff p-value is noted with:

***	pval<0.01
**	0.01<=pval<0.05
*	0.05<= pval < 0.10.

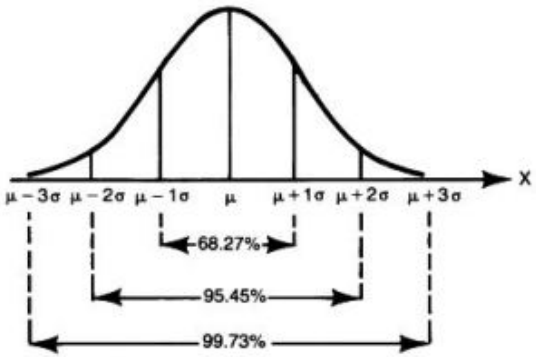
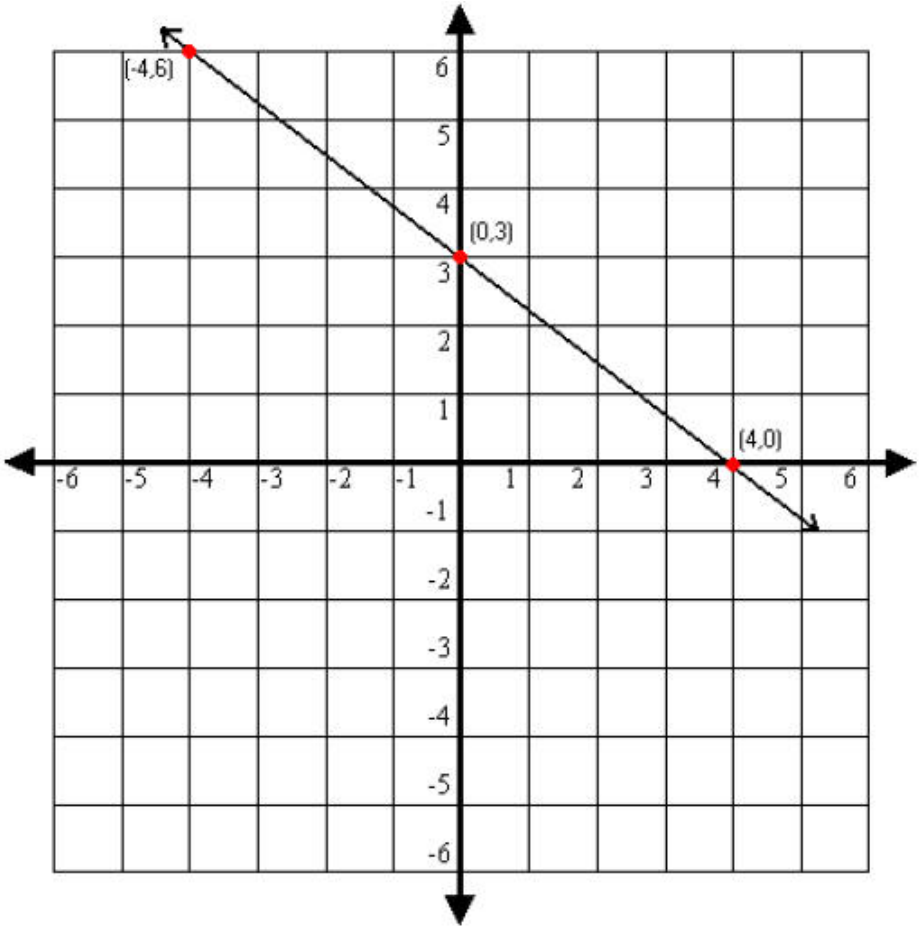


Figure 2

Looking at a graph and identifying the linear equation



- Suppose this is a graph of your patience (y) as a function of traffic jams (x). What is the function?

- Linear equations take the form of y=a+bx. So:
- Step 1: Identify the vertical intercept (0,3) a=3
- Step 2: Identify the horizontal intercept (4,0)
- Step 3: calculate the slope
- increase in y/increase in x

  b = -3/4

  (or rise over run)

- So, function is y=3-3x/4

Inverting a linear function

- traveltime = 14.7 + 0.03*cars
- If it takes you 20 minutes to travel, how many cars are on a freeway?

Inverting a linear function

You know travel time as a function of cars
traveltime = 14.7 + 0.03*cars

You want cars as a function of travel time:
Traveltime- 14.7 = 0.03*cars
Cars = (Traveltime- 14.7) / 0.03
Cars = Traveltime/0.03- 14.7/0.03
Cars = 33.3*Traveltime - 490

Now, it's easier to answer this question:
If it takes you 20 minutes to travel, how many cars are on a freeway?
Cars = 33.3*20 - 490 =176

(BTW: what is the intercept and slope of this inverted function?
Intercept = -490 Slope 33.3)

How many additional businesses should be allowed along a busy highway to maximize citizens satisfaction?

Breaking down the question into mathematical concepts

1. How long does it take to travel the highway? (random variable) On average 26.7 minutes.
2. How does the # of cars affect travel time? (correlation, linear regression, slope) Travel time = 14.7+0.03 cars
3. Can adoption of a different traffic system reduce congestion? (simultaneous equations)
4. How does the # of businesses affect # of cars?(nonlinear equations)
5. What is the optimal # of business to have? (optimization, derivatives, chain rule)

3. comparing two highways: should you adopt another traffic system?

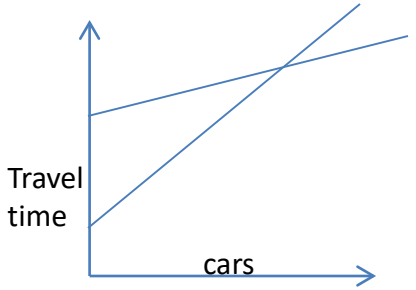
(simultaneous equations, or, systems of equations)

- Previously you learned that for Pittsburgh highways, traveltime = 14.7 + 0.03*cars.

- A colleague suggested that in anticipation of congestion from the new businesses, you should consider a traffic system that has been adopted by Cleveland to reduce travelling time. There, traveltime = 8.7 + 0.05 cars.

- Should you do that? What is the maximum # of cars such that travelling with the Cleveland system is faster than the Pittsburgh system?

- Pittsburgh: Traveltime = 14.7 + 0.03 cars
- Cleveland : Traveltime = 8.7 + 0.05 cars
- The question asks for what is cars such that traveltime is equal to each other.



Several methods:

You can solve a linear systems by:

1. Graphing: draw both lines and see where they meet.
2. Substitution:

Traveltime = 8.7 + 0.05 cars

14.7 + 0.03 cars = 8.7 + 0.05 cars

6 = 0.02 cars.

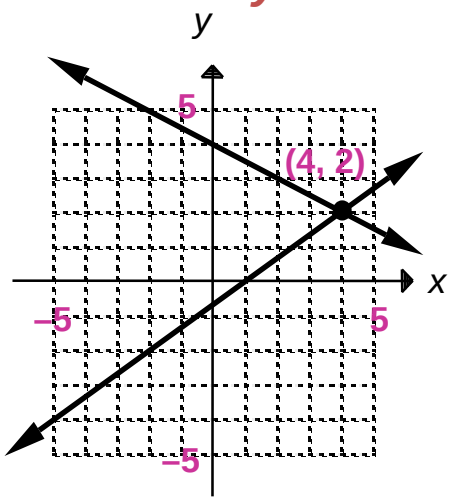
Cars = 300

- Given that mean # of cars on Pittsburgh highways is 385 (see data), the Cleveland system would actually cause more congestion.

Nature of Solutions to Systems of Equations

(A) $2x - 3y = 2$

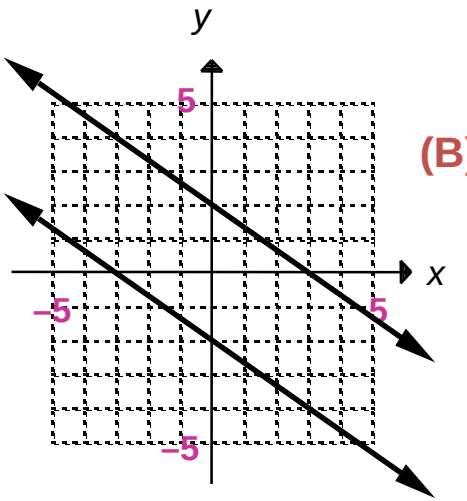
$x + 2y = 8$



Lines intersect at one point only.
Exactly one solution: $x = 4$, $y = 2$

(B) $4x + 6y = 12$

$2x + 3y = -6$

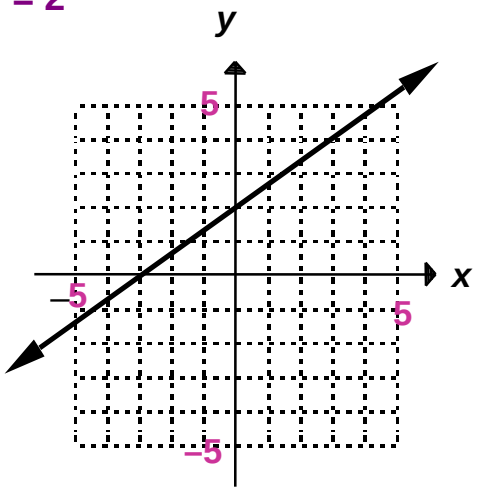


Lines are parallel

No solution.

(C) $2x - 3y = -6$

$-x + {}^{3}/_{2}\, y = 3$



Lines coincide. Infinitely many solutions.

Review Exercise 1

- Questions?

How many additional businesses should be allowed along a busy highway to maximize citizens satisfaction?

Breaking down the question into mathematical concepts

1. How long does it take to travel the highway? (random variable) On average 26.7 minutes.
2. How does the # of cars affect travel time? (correlation, linear regression, slope) Travel time = 14.7+0.03 cars
3. Han adoption of a different traffic system reduce congestion? (simultaneous equations) No.
4. How does the # of businesses affect # of cars?(nonlinear equations)
5. What is the optimal # of business to have? (optimization, derivatives)

4. Nonlinear function

We will now use our other data set, “business.csv”

This data set has the # of businesses on a highway and the # of commuter cars associated with these businesses.

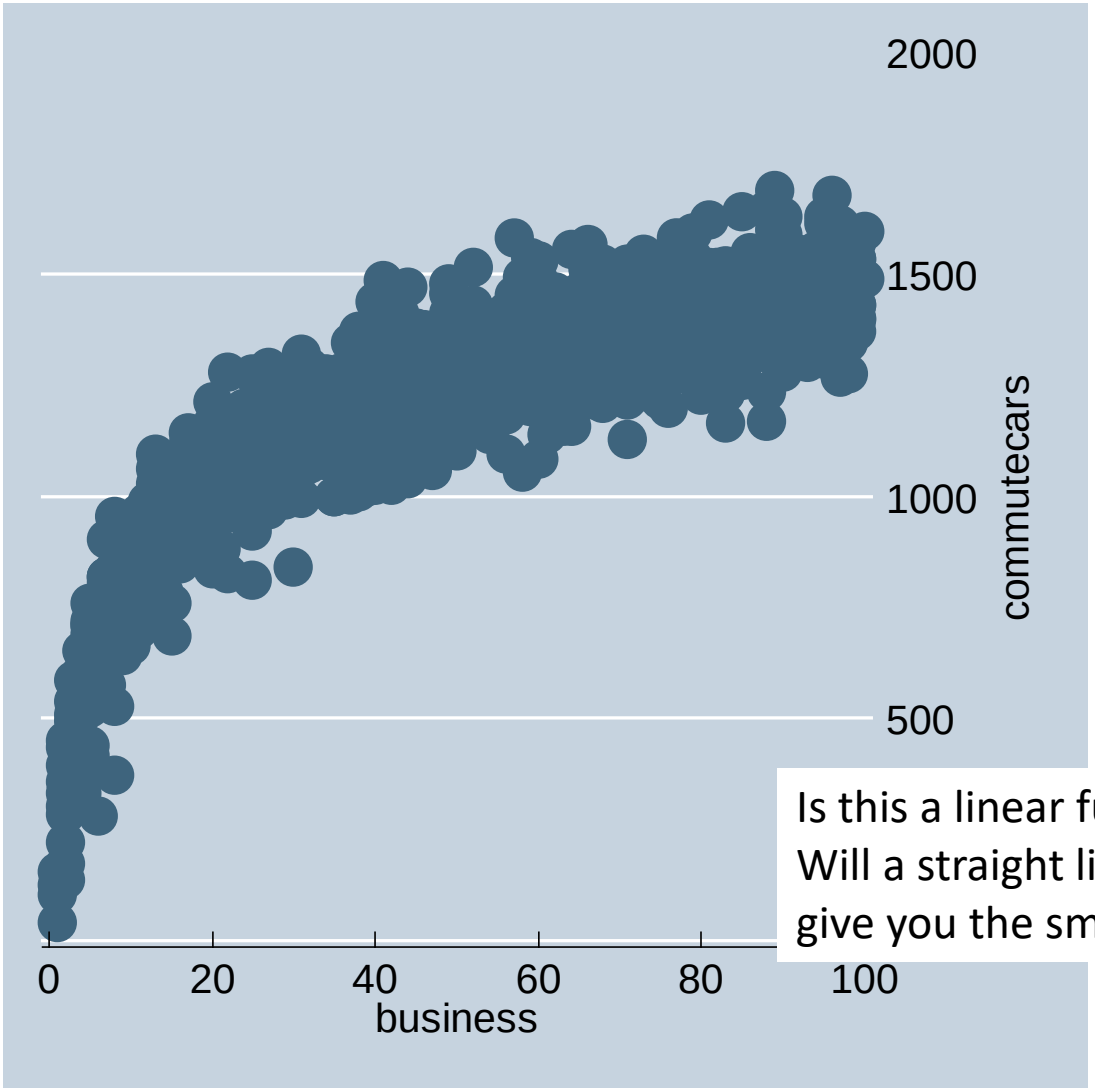
clear (you must clear out the old data)

Load new Business.csv

Look in data editor

What relationship are we trying to figure out?

scatter commutecars business



Is this a linear function?
Will a straight line
give you the smallest error?

Nonlinear functions

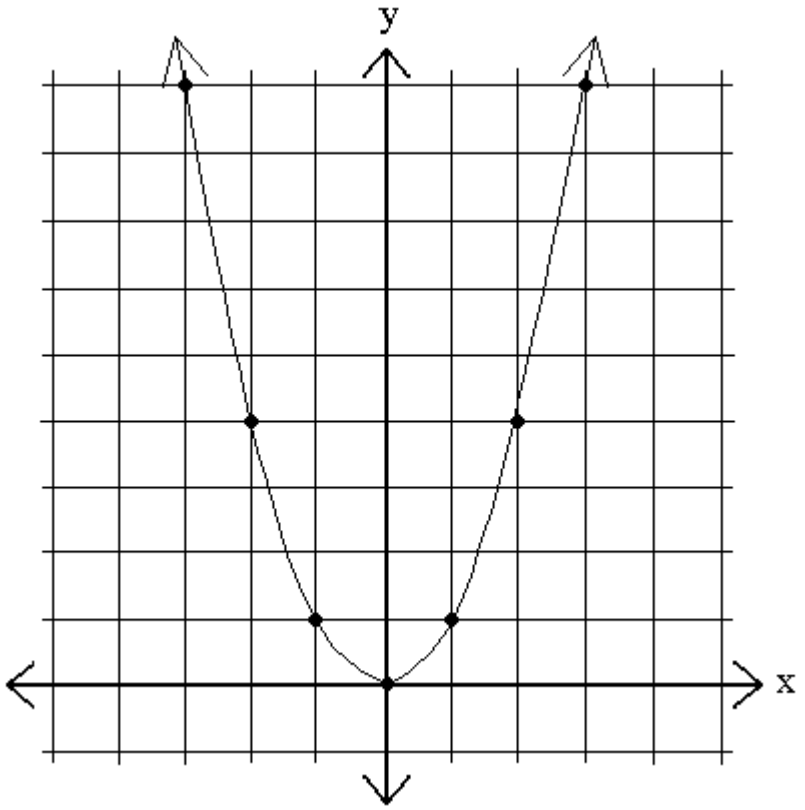
Let’s find what our function resembles:

- Quadratic function
- Logarithmic function
- Exponential function

Quadratic function

$y=x^2$

x	*Y*
-3	9
-2	4
-1	1
0	0
1	1
2	4
3	9

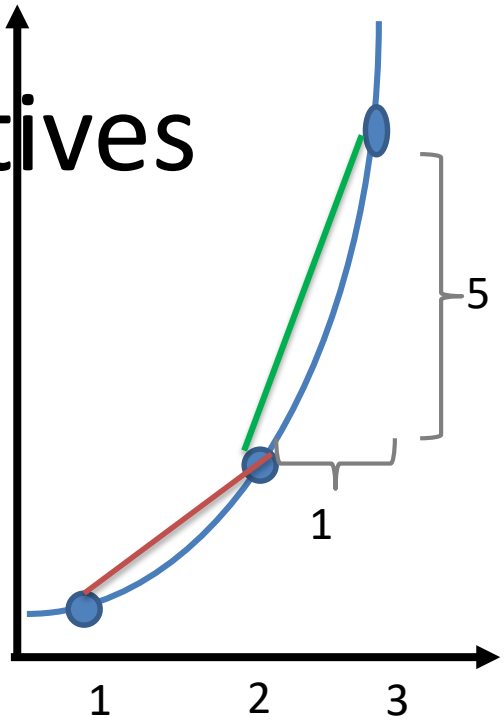


Notice how y changes as x changes.
The slope is no longer the same (“not a constant”) as we travel through the x axis: increasing x by 1 changes y by -5 at x=-3, by 1 at x=0, and by 3 if x=1

Slopes and derivatives

$y=x^2$

x	Y
-3	9
-2	4
-1	1
0	0
1	1
2	4
3	9

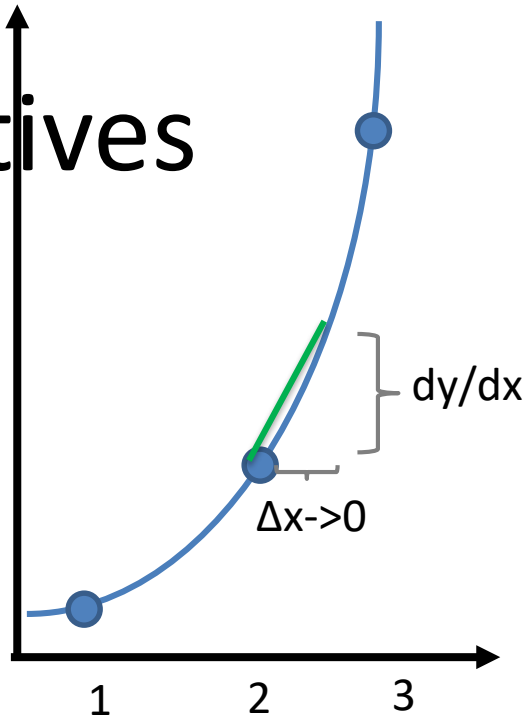


As we discussed earlier, slope is the increase in y / increase in x.
So we can find an average slope between two points.
Average slope as x goes from 2 to 3 is (9-4)/(3-2) = 5
Average slope as x goes from 1 to 2 is (4-1)/(2-1) = 3
But how do we find the slope at the single point (x=2)? There's nothing to measure!

Slopes and derivatives

$y=x^2$

x	Y
-3	9
-2	4
-1	1
0	0
1	1
2	4
3	9



But how do we find the slope at a single point (x) ? We can make up another point x+Δx where Δx is very small, so we have two points (x and x+Δx)and calculate the slope there.

The slope at x **as we make Δx shrink to 0 is the derivative of y at x.**

We write **dx** instead of "as Δx shrink to 0" so the derivative of y over x is usually written as dy/dx.

The Recipe for Derivatives

<u>the power rule</u>:

Identify: m (constant), x (variable), c (exponent)

$$\text{if } y=mx^c \text{ , } dy/dx= mcx^{c-1}$$

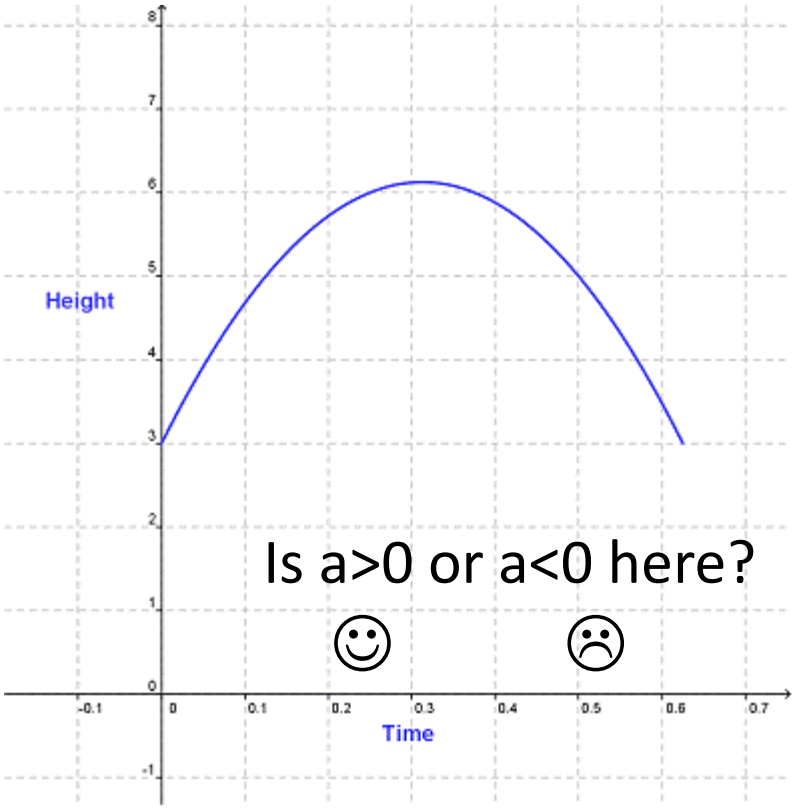
- $y=x^2 = 1x^2$ constant=1, var =x, exponent=2.
  $dy/dx=1*2x^{(2-1)} =2x$

So the derivative of x^2at x=2 is 2*2 = 4

(note this is between 3 and 5 from slide 41)

Other quadratic functions

$y = ax^2 + bx + c$



Quadratic functions are one type of polynomial functions:

$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0,$$

Working with polynomials more generally

Example:

- $Y=3x^8+ 4x^{1/2} - 5x + 2/x+ 9 + 2x^8$

For each term, identify: m constant, x variable, c exponent (mx^c)

- Some special ones:
- $x= x^1$
- $1/x= x^{-1}$ $1/x^2= x^{-2}$
- $1 =x^0$ $x^{1/2} = \text{sqrt}(x)$

- Let's write the polynomial such that we can easily identify the mx^c form:

- $Y=3x^8+ 4x^{1/2} - 5x^1 + 2x^{-1} + 9x^0 +2x^8$

Rules for simplifying polynomials

Product rules	$x^n \cdot x^m = x^{n+m}$	$2^3 \cdot 2^4 = 2^{3+4} = 128$
	$x^n \cdot b^n = (x \cdot b)^n$	$3^2 \cdot 4^2 = (3 \cdot 4)^2 = 144$
Quotient rules	$x^n / x^m = x^{n-m}$	$2^5 / 2^3 = 2^{5-3} = 4$
	$x^n / b^n = (x / b)^n$	$4^3 / 2^3 = (4/2)^3 = 8$
Power rules	$(x^n)^m = x^{n \cdot m}$	$(2^3)^2 = 2^{3 \cdot 2} = 64$

When will you use this in class? When you're working with utility functions.

Derivatives: the "slope" at a point

<u>the power rule</u>:

$$\text{if } y=mx^c \text{ , } dy/dx= mcx^{c-1}$$

- Suppose y is the spread of a disease, x is poverty, and z is temperature, and you're asked to find how poverty affect the spread of a disease.

- $y=3x^3 + 4x^3$. Simplify first: $7x^3$. Then identify constant=7, var =x, exponent=3.
  $dy/dx=7*3x^{(3-1)} =21x^2$
- $y=3x^2+ 8$ constant=8, var =x, exponent=0.
  $dy/dx=6x + 8*0x^{(0-1)} =6x + 0 = 6x$
- $y=3x^2+ 8z$ constant=8z, var =x, exponent=0.
  $dy/dx=6x + 8z*0x^{(0-1)} =6x + 0 = 6x$
- $y=x^{-1} + 8z$ constant=1, var=x, exponent=-1.
  $dy/dx=1*-1x^{(-1-1)} + 0 = -1x^{-2} = -1/x^2$

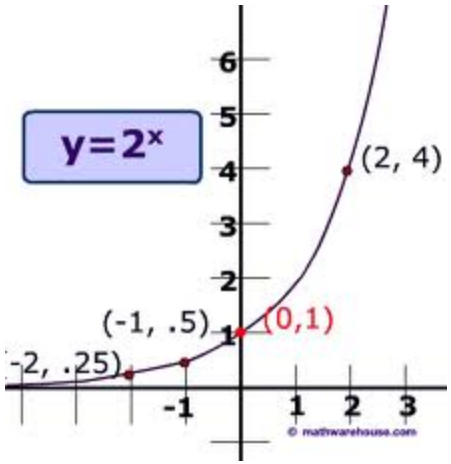
When will you use this in class? When you're trying to figure out the rate of change in an outcome due to the implementation of a policy.

Exponential function

- The growth of a terrorist cell:
- At month 0 there's 1 person 1
- At month 1 this person recruited 2 people 2
- At month 2 each persons recruited 2 people 4
- What is the function that describe the growth?
- $f=2^x$ where x is time (month)

This is an exponential functions

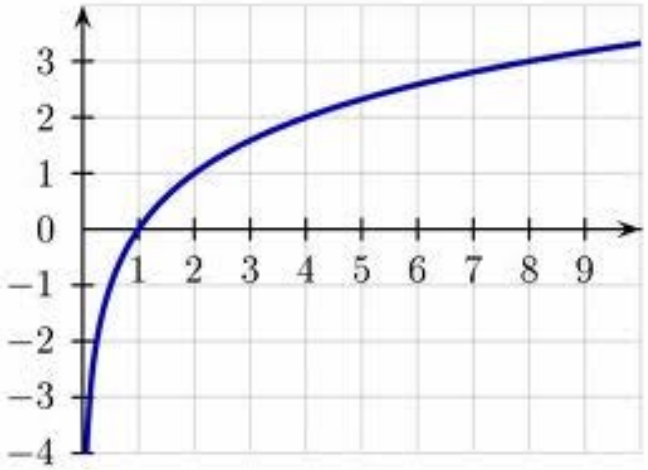
Notice it "asymptotes" at the y axis.



Logarithmic function

- Time since the inception of the terrorist cell
- If there is 1 member it must have just started t=0
- If there are 2 members it must have been last month. t=1
- If there are 32 members t =?
  (5 months)
- Equation: $y=\log_2 x$ where x is # of members and y is months
- "$\log_a x$" means "to what power (exponent) must a be raised to get x?

- This is the inverse
  of the exponential function
- Notice it “asymptotes” at the x axis.



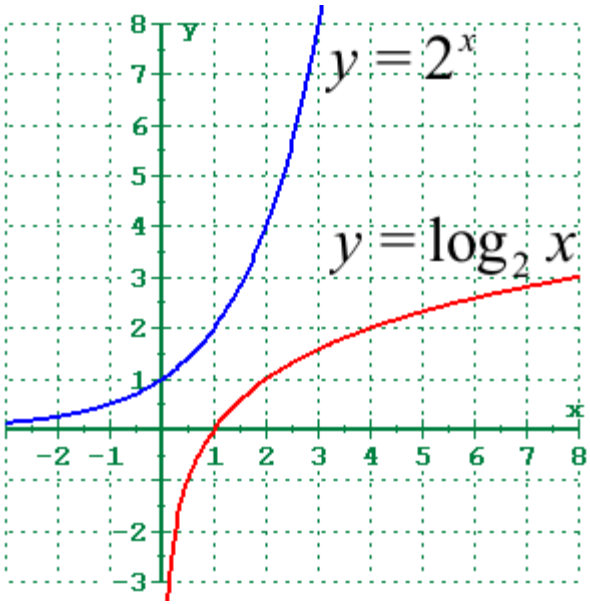
Some rules for dealing with logs

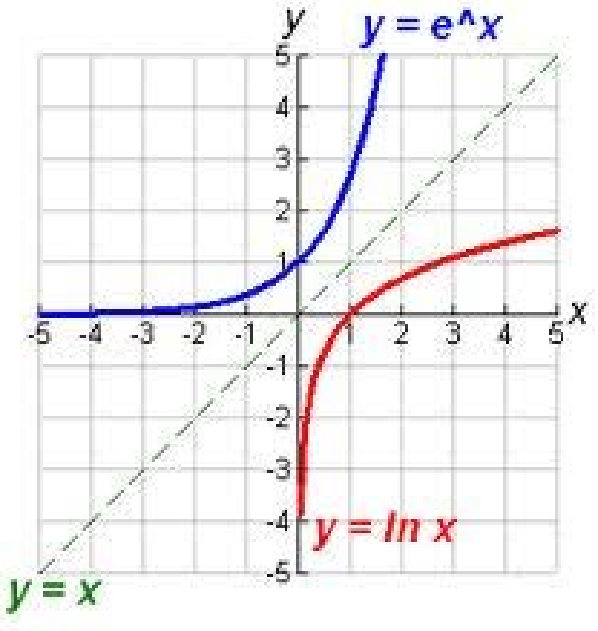
- $\log_2 2 = 1$ since $2^1 = 2$
- $\log_2 4 = 2$ since $2^2 = 4$
- $\log_2 32 = \log_2 2*16 = \log_2 2 + \log_2 16$

  since $2^1 * 2^4 = 2^5 = 16$
- $\log_2 1/16 = \log_2 1 - \log_2 16$

  since $2^0 / 2^4 = 2^{-4}$

2^x and exp(x)
Logs and natural logs





Exp(x) = 2.72^x

This quantity is often used when quantities grow proportionally to it's value. It is often seen in math, physics, and chemistry.

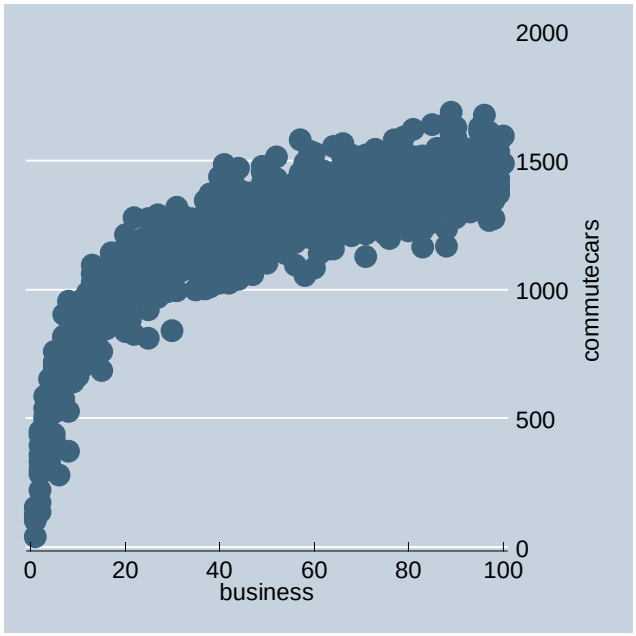
When will you use this? When you're learning about logistic regressions.

We will mostly work with natural log (ln), because their derivatives are easier.

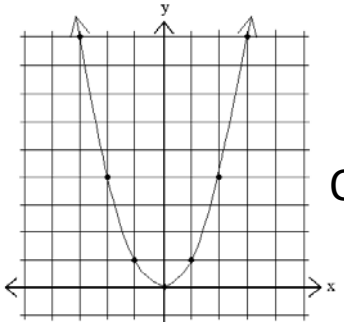
- $\ln e = 1.$ since e^1 = e
- $\ln ab = \ln a + \ln b.$
- $\ln a / b = \ln a - \ln b.$
- $\ln a^n = n \ln a.$

- Compare:
- y=log a (x) dy/dx= 1/ (x ln a)
- With:
- y=ln (x) dy/dx= 1/ x
- Also,
- $y=e^x$ $dy/dx=e^x$

So back to your data:

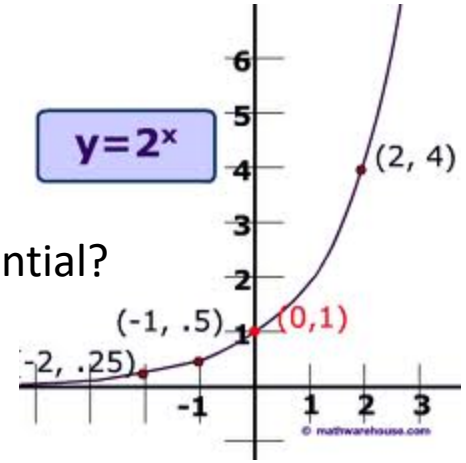


Indeed, cars = 130+290*ln(business)

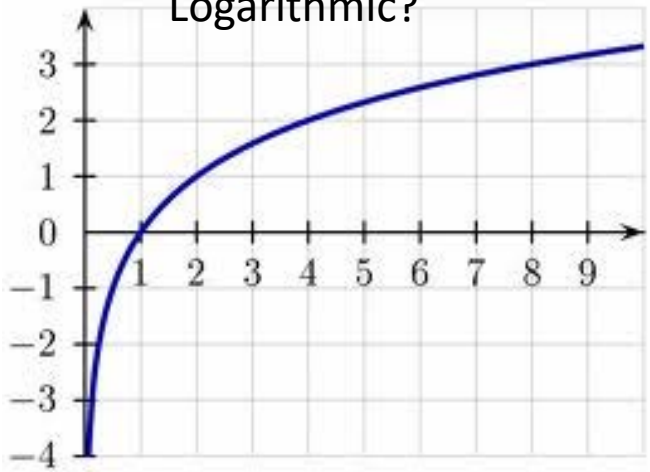


Quadratic?

Exponential?

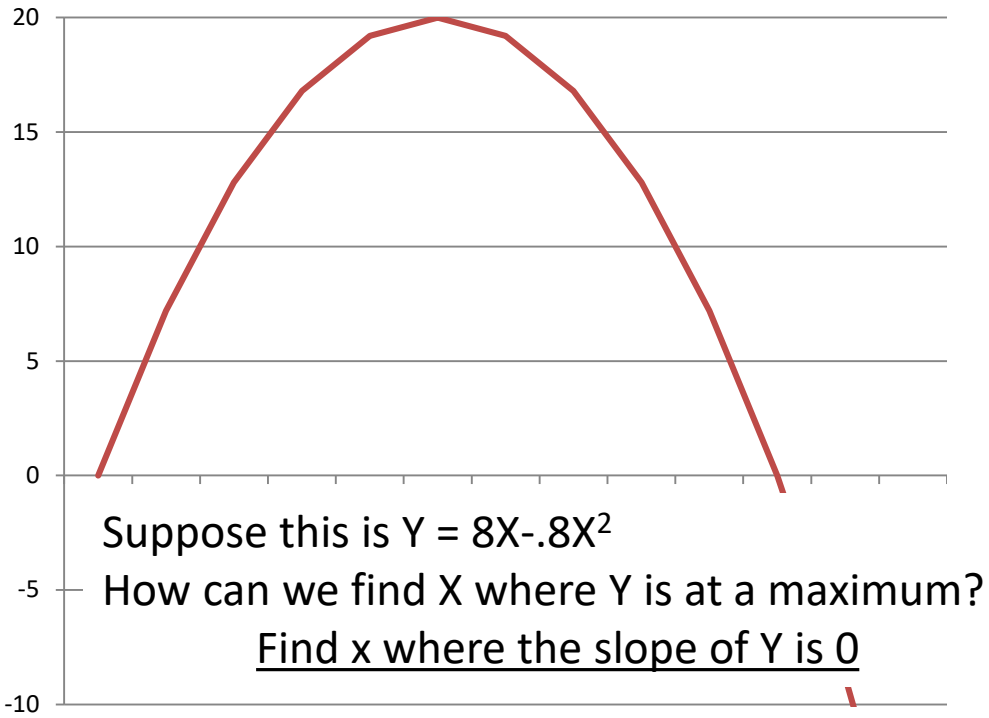


Logarithmic?



5. what is the optimal # of business to have? (optimization, compound functions)

How do we optimize a function?



General Optimization

Recall power rule: if $Y=mx^c$, $dY/dX= mcx^{c-1}$

Supposed want to maximize a function:

$Y = 8X-.8X^2$

First we have to take a derivative:

$dY/dx = 8*1X^{1-1}-.8*2X^{2-1}$

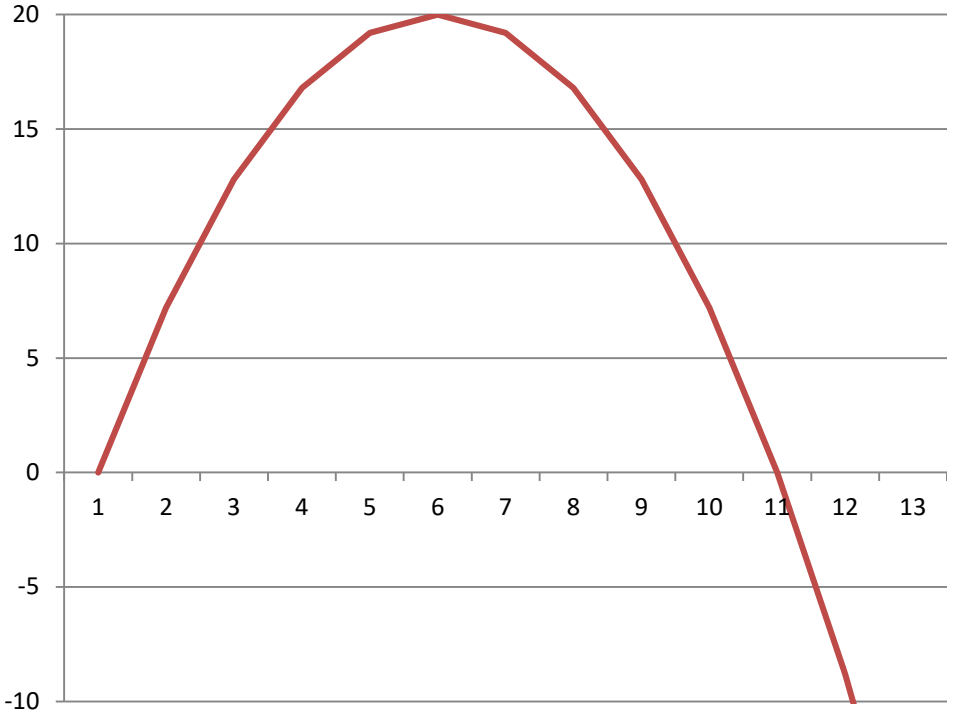
$= 8-1.6X$

Then when we set it to 0, we can solve for X that maximizes the function

$8-1.6X = 0$

$8 = 1.6X$

$X=5$



$$Y = 8X-0.8X^2$$

$$dY/dX = 8 – 0.8*2*X = 8-1.6X$$

X	Y
0	0
1	7.2
2	12.8
3	16.8
4	19.2
5	20
6	19.2
7	16.8
8	12.8
9	7.2
10	0
11	-8.8
12	-19.2

How many businesses should be on the highway?

Let's break this down:

1. How does business affect travel time?

   We know how cars affect travel time

   And we know how businesses affect cars.

2. Suppose his public opinion expert says:

complaints = travel time,

praise = # of business^2/2

Then how can he maximize:

praise – complaints ?

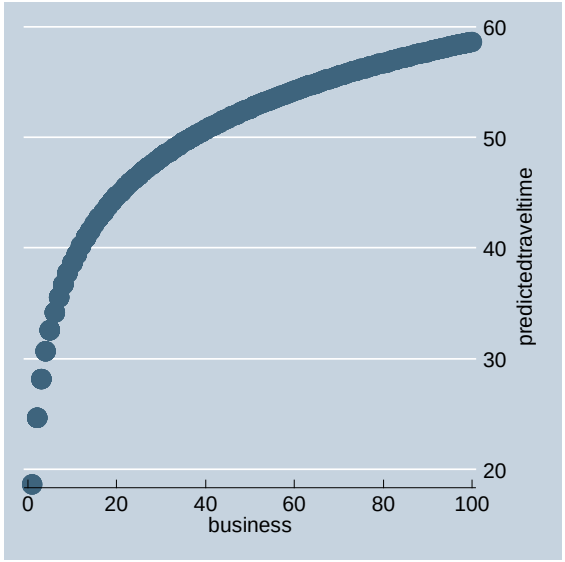
You have cars and traveltime.

- cars = 130+290*ln(business)

You have business and cars.

- traveltime = 14.7 + 0.03*cars

How do you combine them?



- traveltime = 14.7 + 0.03* (130+290*ln(business))
- traveltime = 18.6+8.7ln(business)

Now we are ready to use the info from the public opinions guy:

- Praise = # of business^2/2
- Complaints = traveltime = 18.6+8.7ln(business)

- We want to maximize: Benefit= Praise – Complaints
- Benefit = business^2/2 -18.6-8.7ln(business)

- Take derivative:
- d Benefit / d business = business – 8.7/business
- Set to 0, we get:
- 0 = business – 8.7/business
- 8.7/business = business
- 8.7 = business^2
- Optimal number of business = sqrt(8.7) = 2.95 or 3 businesses

Exercise 2

- Break

Review Exercise 2

- Any questions?

Writing and saving commands in STATA

- In your classes (and in your job in the future) you will want more control over what you did to the data and replicability.
- This is so you can remember what you did and that others can replicate your results.
- This is harder to do with the menu bar.

  - Go to Window, Do File Editor, and choose New Do-file Editor.
  - This will open a new .do file.
  - Write your commands in it.
  - Highlight one of the commands and click the “Execute (do)” icon. It should run the command. You can also copy and paste directly to the command window.
  - Save this file as MathCamp.do
  - Continue adding commands into this file.

Loading and exploring

- Clearing memory: clear
- Loading .csv file: cars.csv
- See all variables: sum
- String variables (highway) vs numeric variables
- Tab highway

Variable	Obs	Mean	Std. Dev.	Min	Max
cars	1,674	385.3883	232.479	0	1230
traveltime	1,674	26.71808	8.605933	11.16805	61.63294
highway	0				
ID	1,674	8.378136	4.848058	0	17

.

Relationship between variables

- Pwcorr (correlation)

```
. pwcorr cars traveltime

             |     cars travel~e
-------------+------------------
        cars |   1.0000
  traveltime |   0.8414   1.0000
```

- Reg traveltime cars
- Scatter traveltime cars

Sorting and Viewing Data

- gsort
- Sort in both order
- Ascending: gsort cars
- Descending: gsort -cars
- Sort
- Only sort in ascending order

- List

.

gsort cars

. list in 1/5

```
+---------------------------------+
v1   cars   travel~e   highway
---------------------------------
1.  1153     0   18.31223    Roscoe
2.  1532     0   15.03788    Roscoe
3.  1321     0   15.07491    Roscoe
4.   170     0   18.04261     Robb
5.   822     0    12.8783   Jemison
+---------------------------------+
```

. list in -5/L

```
+---------------------------------+
v1   cars   travel~e   highway
---------------------------------
1670.   220   1170   57.35289     Robb
1671.   253   1190   45.25637     Robb
1672.   554   1210   55.85953   Clarion
1673.  1390   1220   54.14872    Roscoe
1674.    59   1230   47.82588    Roscoe
+---------------------------------+
```

Conditional statements and working with strings (if, and (&), or (|), ==, !=)

sum traveltime if cars < 100

mean traveltime if cars > 150 & cars < 200

reg traveltime cars if highway !="SqHill"

sum traveltime if highway =="SqHill" | highway == "Clarion"

list traveltime if highway =="SqHill" & cars > 400

Generating new variables

- gen: simple transformations of other variables

  gen travelsq = traveltime^2

- What if you mess up making a variable and want to recreate it? Eg. You want travelsq to be ½*traveltime^2

  drop travelsq

  gen travelsq = (1/2)* traveltime^2

Can combine gen with logical statements :
gen toocrowded = (cars>400)

Using your new variable:
reg traveltime cars if toocrowded
reg traveltime cars if !toocrowded
reg traveltime toocrowded

Graphing

Comparing two subgroups:

twoway (scatter traveltime cars if toocrowded) (scatter traveltime cars if !toocrowded)

twoway (scatter traveltime cars if highway==“Roscoe”) (scatter traveltime cars if highway==“Robb”)

Comparing two version of traveltime:

twoway (scatter traveltime cars) (scatter travelsq cars)

How to save your graphs?

File– Save As – (I usually do .pdf)

Or: Win users: right click and click Copy and then paste into your word doc.

Review Exercise 3

- Any questions?

On to the Race!

- Show Race Packet Materials.
- Tomorrow: you will absolutely need your computer.
- You will be coding and thinking and racing from room to room, so make sure you are comfortable.
- There will be 10 **clues**. Solving each clue in three tries or less will earn your team 1 point. The team with the highest number of points wins the race. **Ties** are broken by how quickly you complete the race.
- There will be Roadblocks. In Roadblocks each person in the team must solve a puzzle individually. The point will only be given if both team members successfully solve their puzzle.

How to win?

- Review all the material tonight with your teammate and decide on how you want to handle roadblocks and other scenarios. The math will be simple but will require creative applications.
- Stata commands: You MUST get familiar with all the commands we did today.
- When getting your answers checked you can send just one person so one of you can continue working.
- Tomorrow: you can setup starting from 12pm. We will distribute materials for the race at 1pm
- On to work with your teammate! (Group exercise)

Training done!