Week 1, video 2:

Regressors

Prediction

- Develop a model which can infer a single aspect of the data (predicted variable) from some combination of other aspects of the data (predictor variables)
- Sometimes used to predict the future
- Sometimes used to make inferences about the present

Prediction: Examples

- A student is watching a video in a MOOC right now.
  - **Is he bored or frustrated?**
- A student has used educational software for the last half hour.
  - **How likely is it that she knows the skill in the next problem?**
- A student has completed three years of high school.
  - **What will be her score on the college entrance exam?**

What can we use this for?

- Improved educational design
  - If we know when students get bored, we can improve that content
- Automated decisions by software
  - If we know that a student is frustrated, let's offer the student some online help
- Informing teachers, instructors, and other stakeholders
  - If we know that a student is frustrated, let's tell their teacher

Regression in Prediction

- There is something you want to predict ("the label")
- The thing you want to predict is numerical
  - Number of hints student requests
  - How long student takes to answer
  - How much of the video the student will watch
  - What will the student's test score be

Regression in Prediction

- A model that predicts a number is called a regressor in data mining
- The overall task is called regression

Regression

- To build a regression model, you obtain a data set where you already know the answer – called the *training label*
- For example, if you want to predict the number of hints the student requests, each value of numhints is a training label

Skill	pknow	time	totalactions	numhints
ENTERINGGIVEN	0.704	9	1	0
ENTERINGGIVEN	0.502	10	2	0
USEDIFFNUM	0.049	6	1	3
ENTERINGGIVEN	0.967	7	3	0
REMOVECOEFF	0.792	16	1	1
REMOVECOEFF	0.792	13	2	0
USEDIFFNUM	0.073	5	2	0
….				

Regression

- Associated with each label are a set of “features”, other variables, which you will try to use to predict the label

Skill	pknow	time	totalactions	numhints
ENTERINGGIVEN	0.704	9	1	0
ENTERINGGIVEN	0.502	10	2	0
USEDIFFNUM	0.049	6	1	3
ENTERINGGIVEN	0.967	7	3	0
REMOVECOEFF	0.792	16	1	1
REMOVECOEFF	0.792	13	2	0
USEDIFFNUM	0.073	5	2	0
….				

Regression

- The basic idea of regression is to determine which features, in which combination, can predict the label's value

Skill	pknow	time	totalactions	numhints
ENTERINGGIVEN	0.704	9	1	0
ENTERINGGIVEN	0.502	10	2	0
USEDIFFNUM	0.049	6	1	3
ENTERINGGIVEN	0.967	7	3	0
REMOVECOEFF	0.792	16	1	1
REMOVECOEFF	0.792	13	2	0
USEDIFFNUM	0.073	5	2	0
….				

Linear Regression

- The most classic form of regression is linear regression

- Numhints = 0.12*Pknow + 0.932*Time – 0.11*Totalactions

Skill	pknow	time	totalactions	numhints
COMPUTESLOPE	0.544	9	1	?

Quiz

Skill	pknow	time	totalactions	numhints
COMPUTESLOPE	0.322	15	4	?

- Numhints = 0.12*Pknow + 0.932*Time – 0.11*Totalactions
- What is the value of numhints?

A) 8.34
B) 13.58
C) 3.67
D) 9.21
E) FNORD

Quiz

- Numhints = 0.12*Pknow + 0.932*Time – 0.11*Totalactions

- Which of the variables has the largest impact on numhints?
  (Assume they are scaled the same)

A) Pknow
B) Time
C) Totalactions
D) Numhints
E) They are equal

However…

- These variables are unlikely to be scaled the same!
- If Pknow is a probability
  - From 0 to 1
  - We'll discuss this variable later in the class
- And time is a number of seconds to respond
  - From 0 to infinity
- Then you can't interpret the weights in a straightforward fashion
  - You need to transform them first

Transform

- When you make a new variable by applying some mathematical function to the previous variable
- $Xt = X^2$

Transform: Unitization

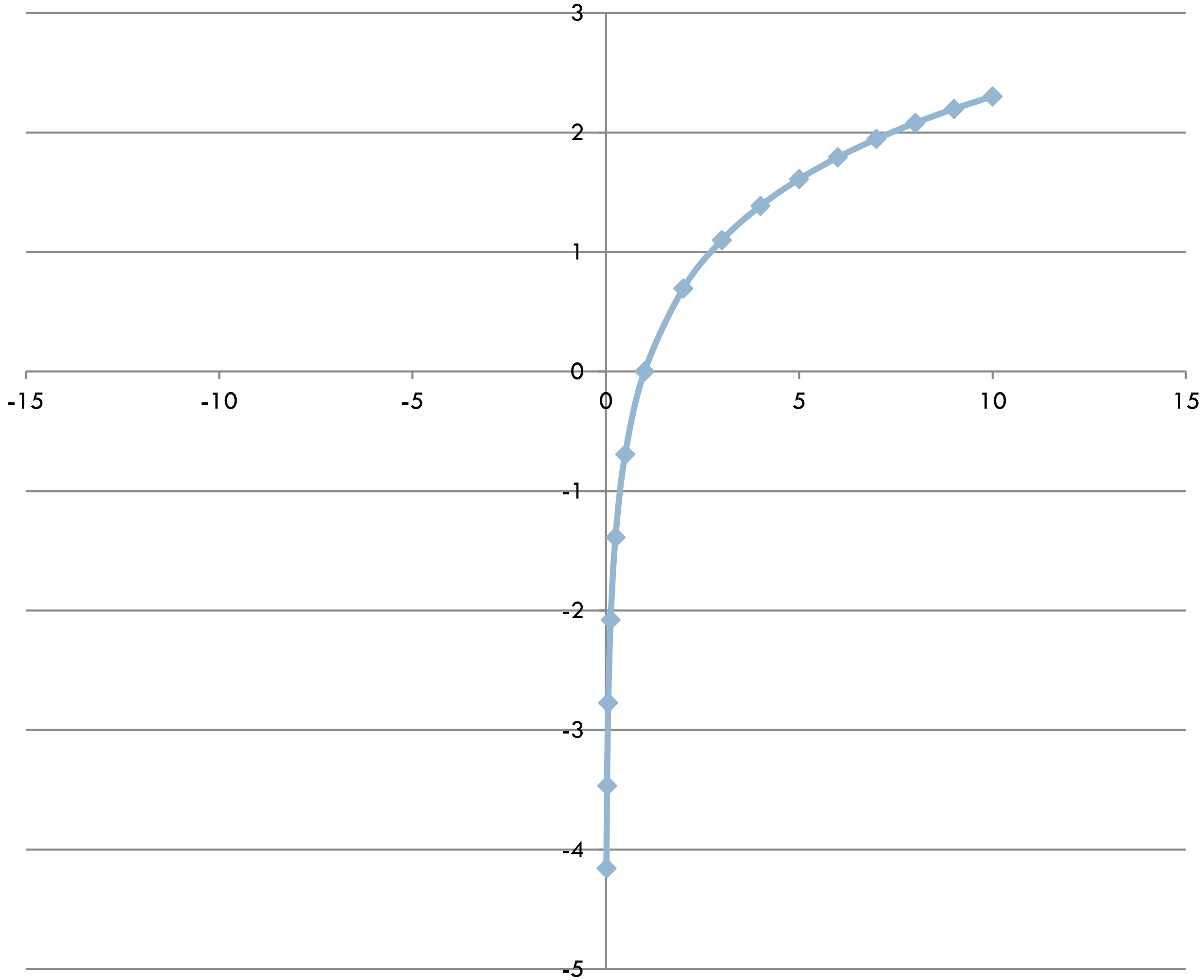
- Increases interpretability of relative strength of features
- Reduces interpretability of individual features

$$Xt = \frac{X - M(X)}{SD(X)}$$

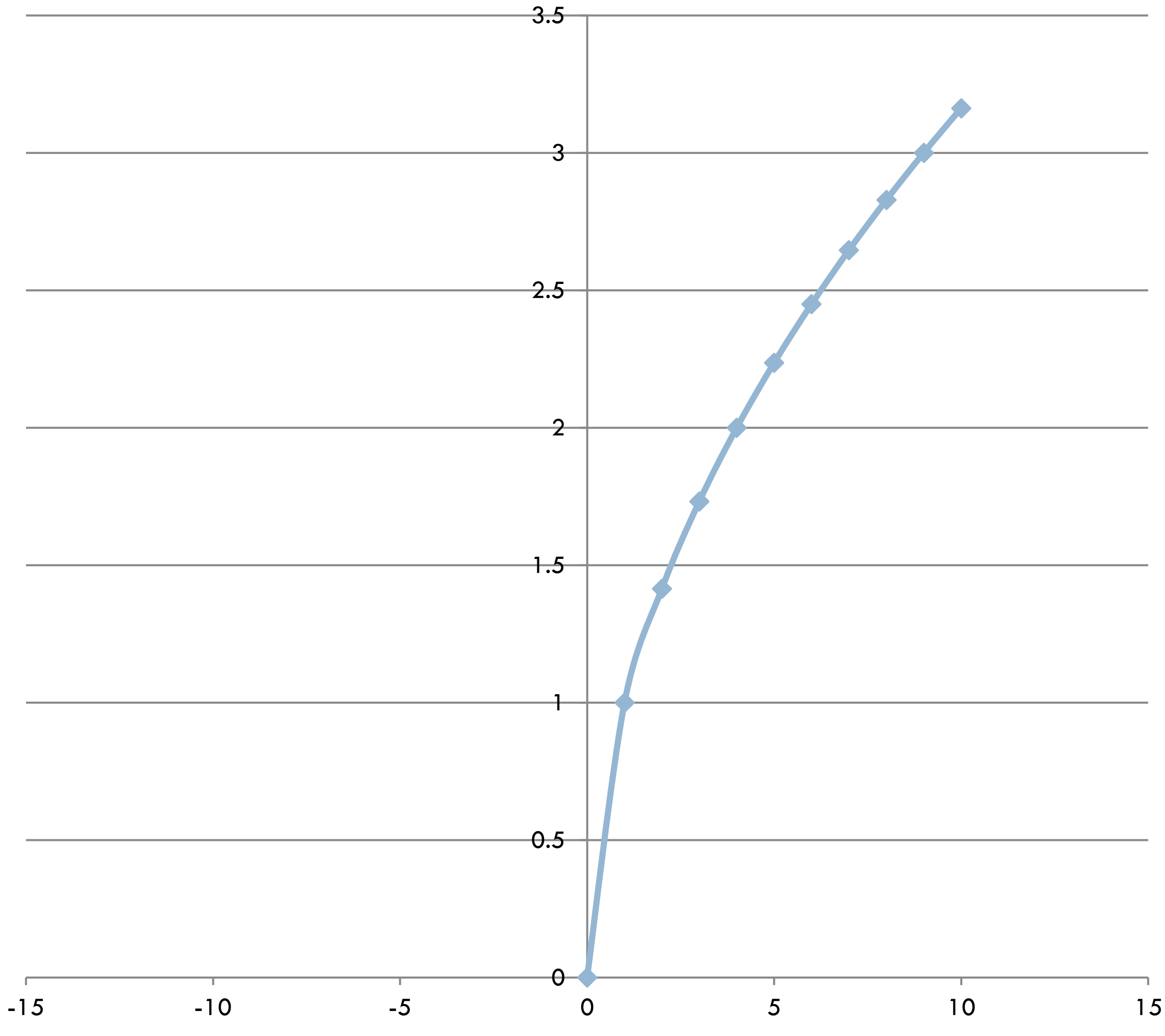
Linear Regression

- Linear regression only fits linear functions…
- Except when you apply transforms to the input variables
- Which most statistics and data mining packages can do for you

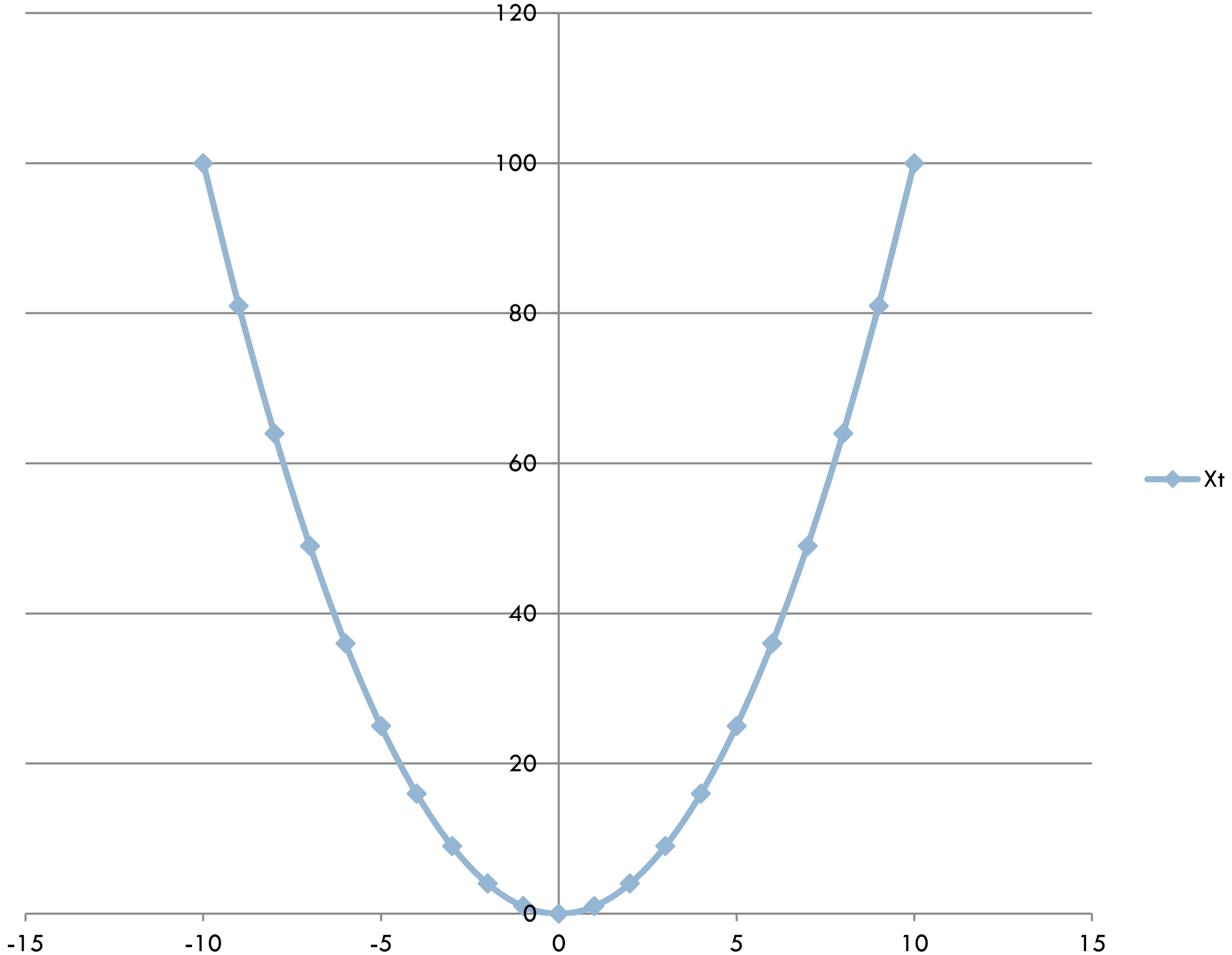
Ln(X)



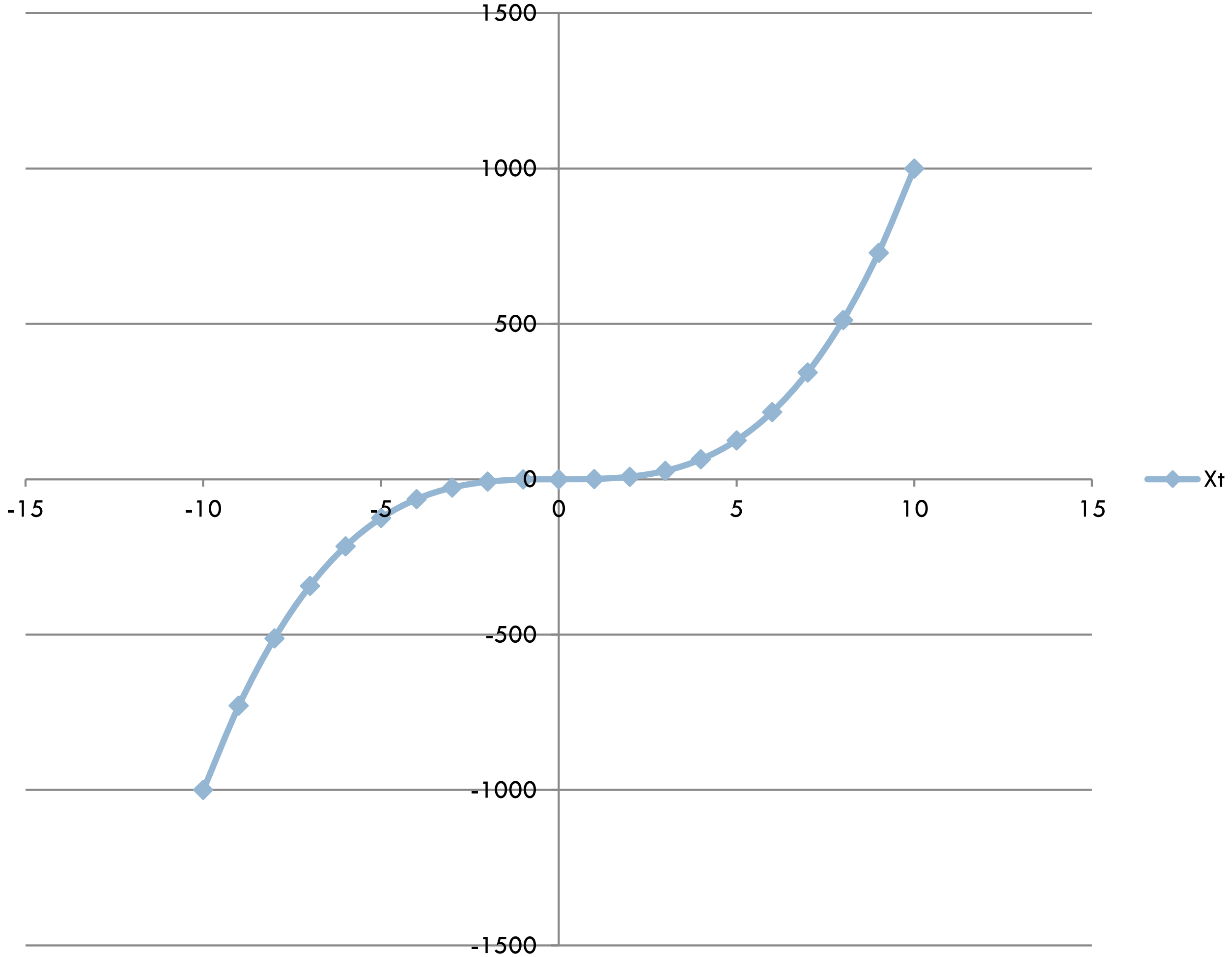
Sqrt(X)



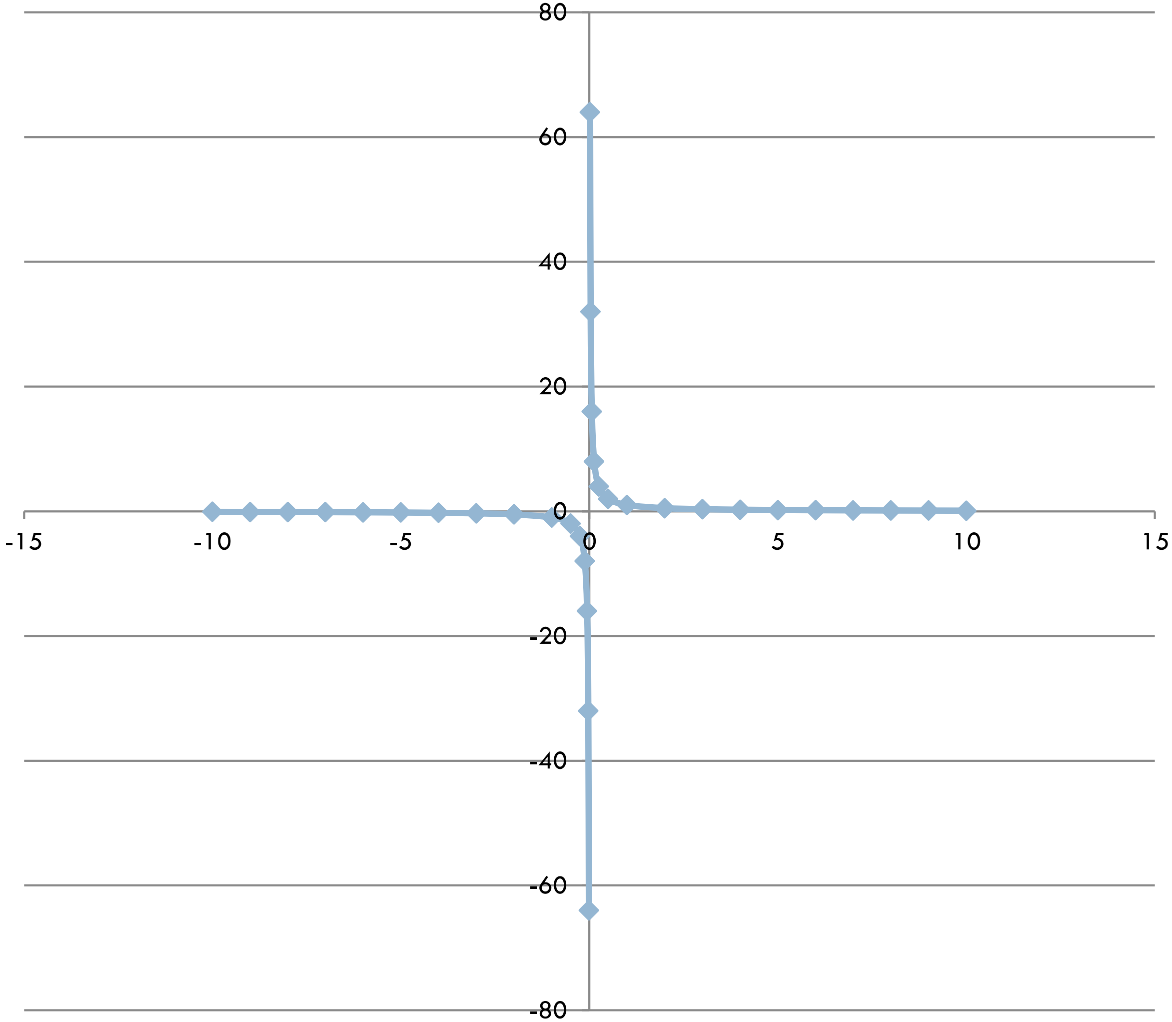
X^2



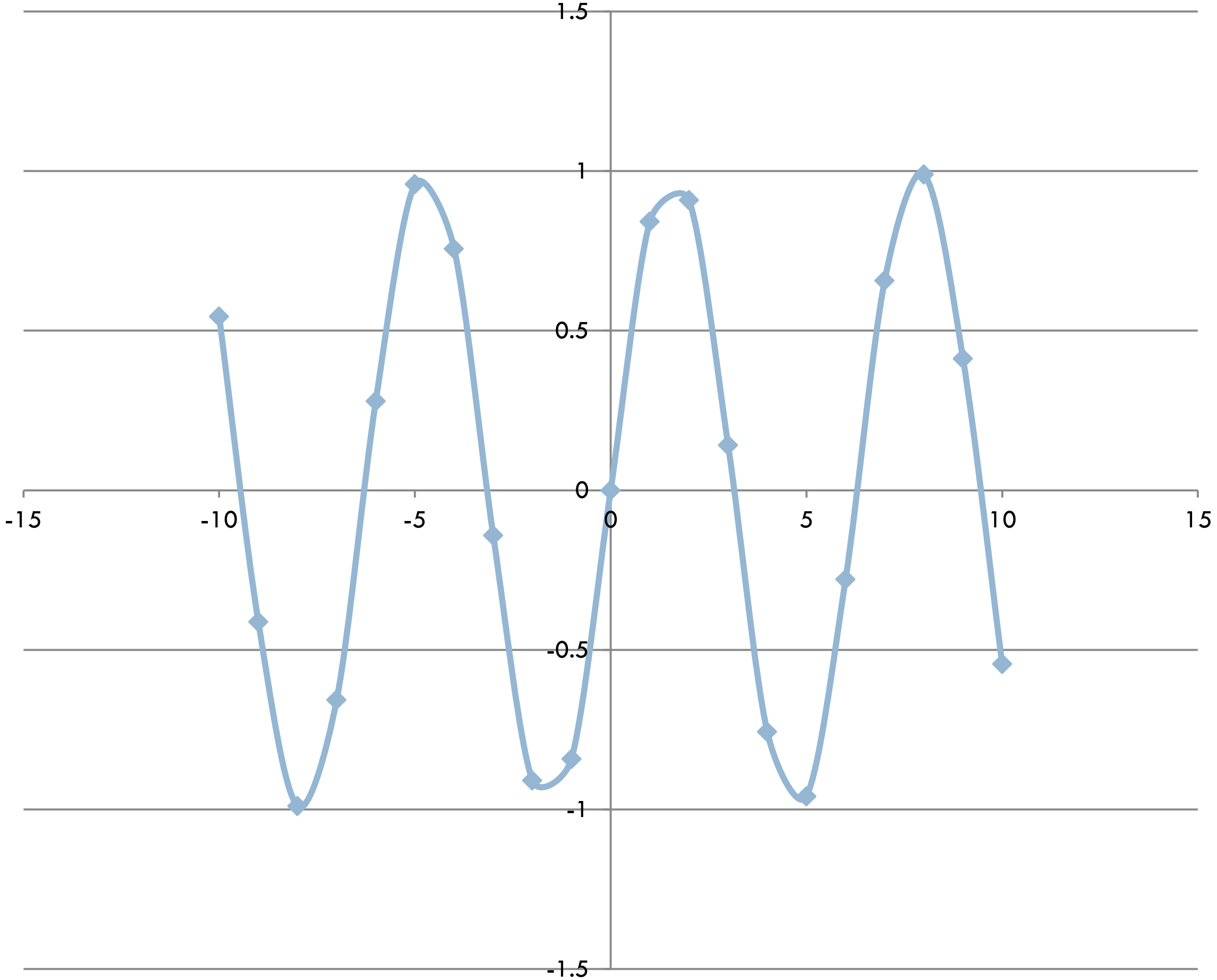
X^3



1/X



Sin(X)



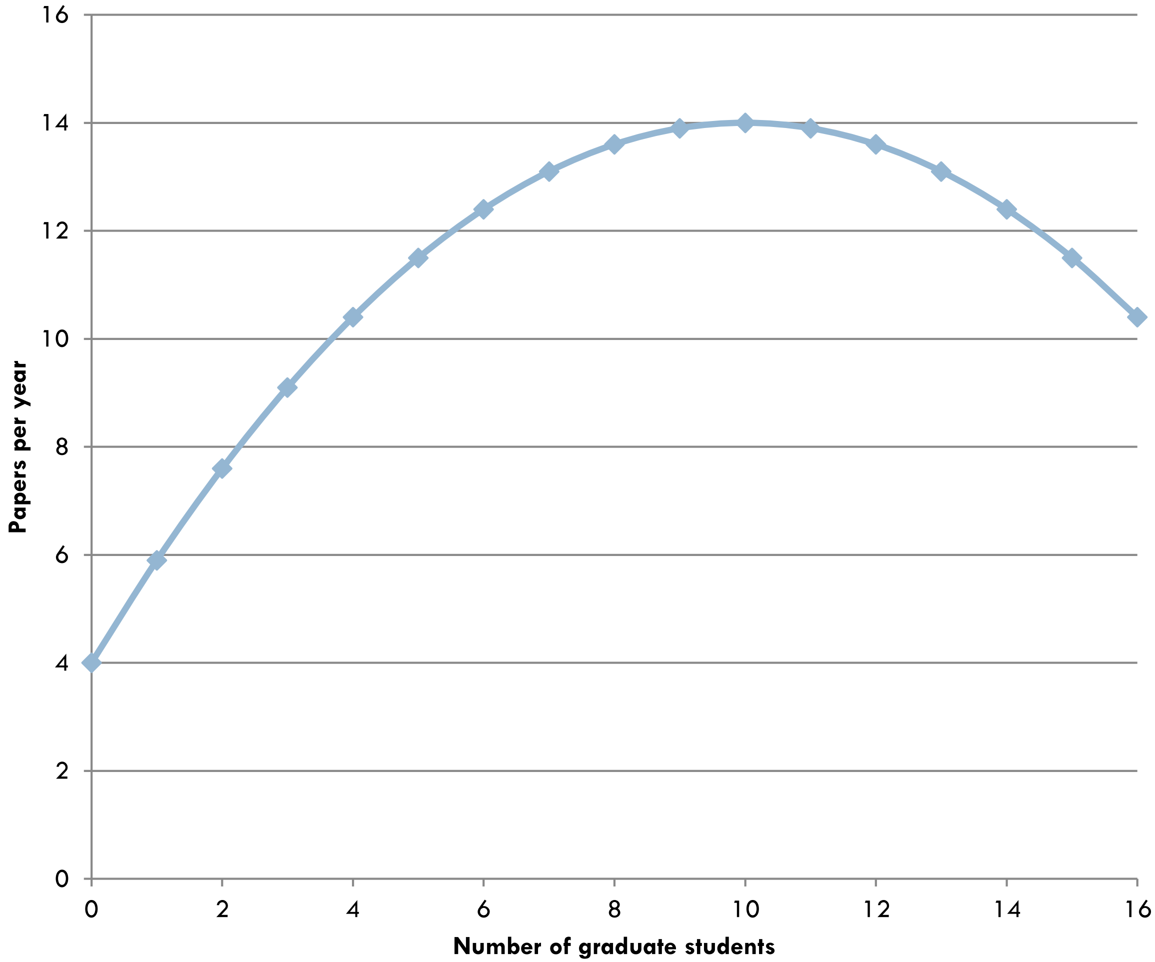
Linear Regression

- Surprisingly flexible…
- But even without that
- It is blazing fast
- It is often more accurate than more complex models, particularly once you cross-validate
  - Caruana & Niculescu-Mizil (2006)
- It is feasible to understand your model (with the caveat that the second feature in your model is in the context of the first feature, and so on)

Example of Caveat

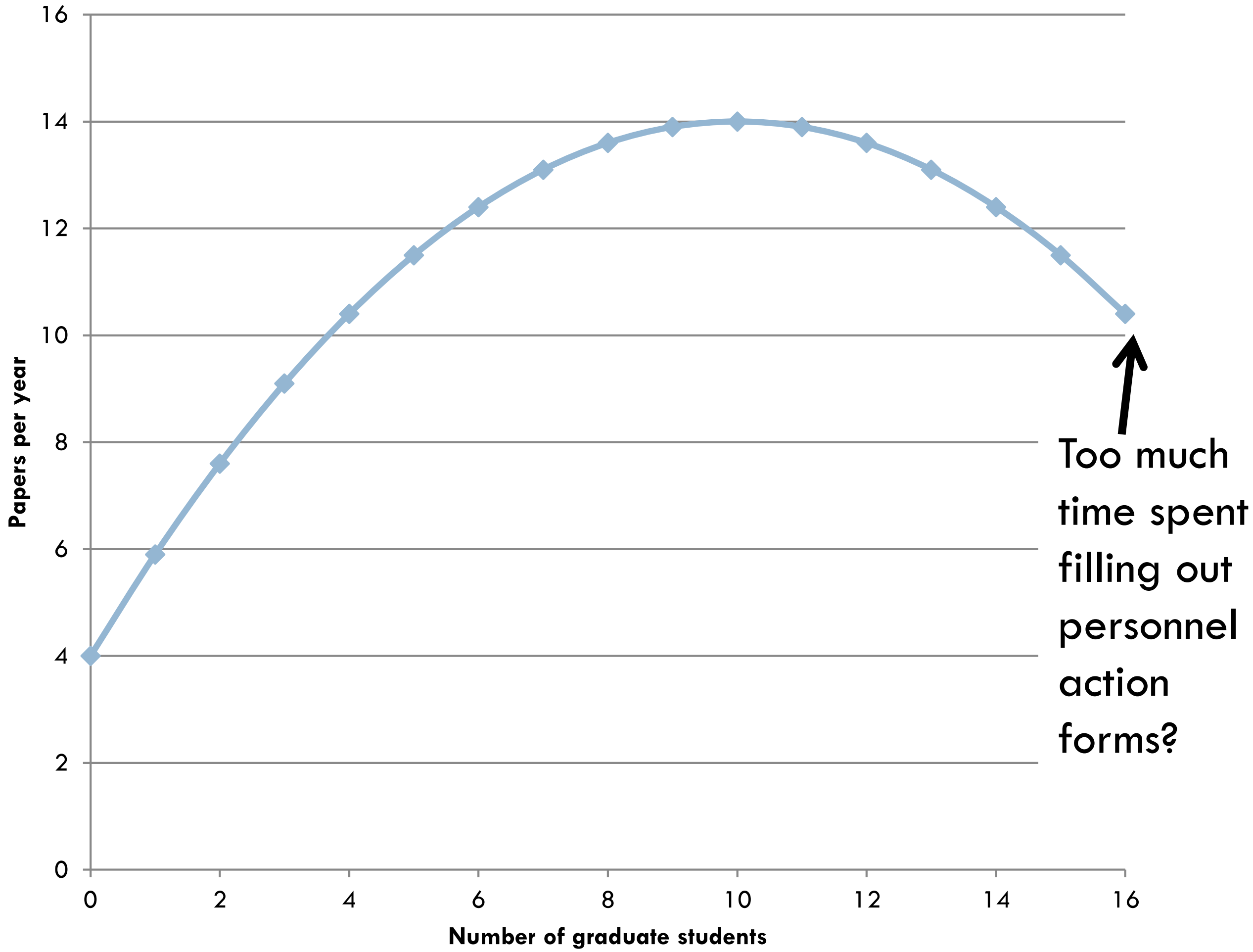
- Let's graph the relationship between number of graduate students and number of papers per year

Data



16
14
12
10
8
6
4
2
0
Papers per year
0
2
4
6
8
10
12
14
16
Number of graduate students

Data

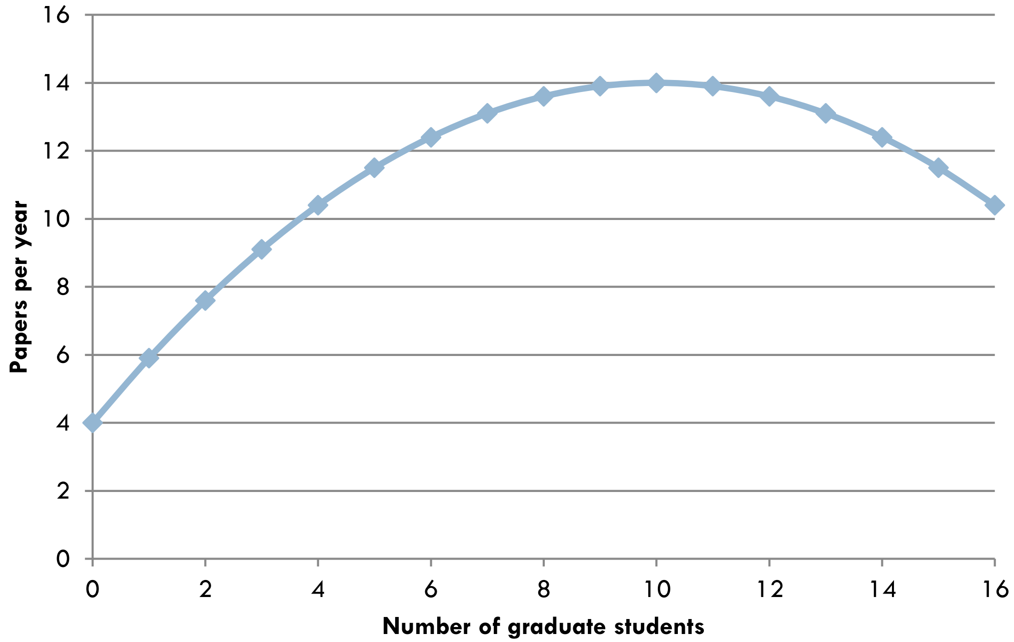


Papers per year
Number of graduate students
16
14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14
16
Too much time spent filling out personnel action forms?

Model

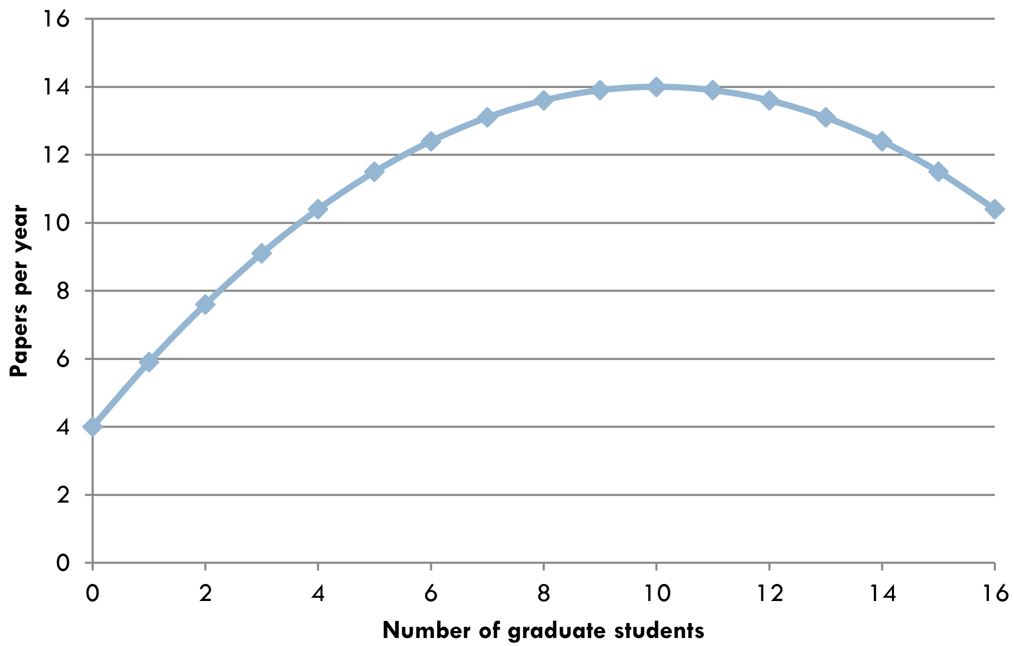
- Number of papers =
  4 +
  2 * # of grad students
  $- 0.1 * (\# \text{ of grad students})^2$
- But does that actually mean that $(\# \text{ of grad students})^2$ is associated with less publication?
- No!

Example of Caveat



- (# of grad students)^2 is actually positively correlated with publications!
  - r=0.46

Example of Caveat



- The relationship is only in the negative direction when the number of graduate students is already in the model…

Example of Caveat

- So be careful when interpreting linear regression models (or almost any other type of model)

Regression Trees

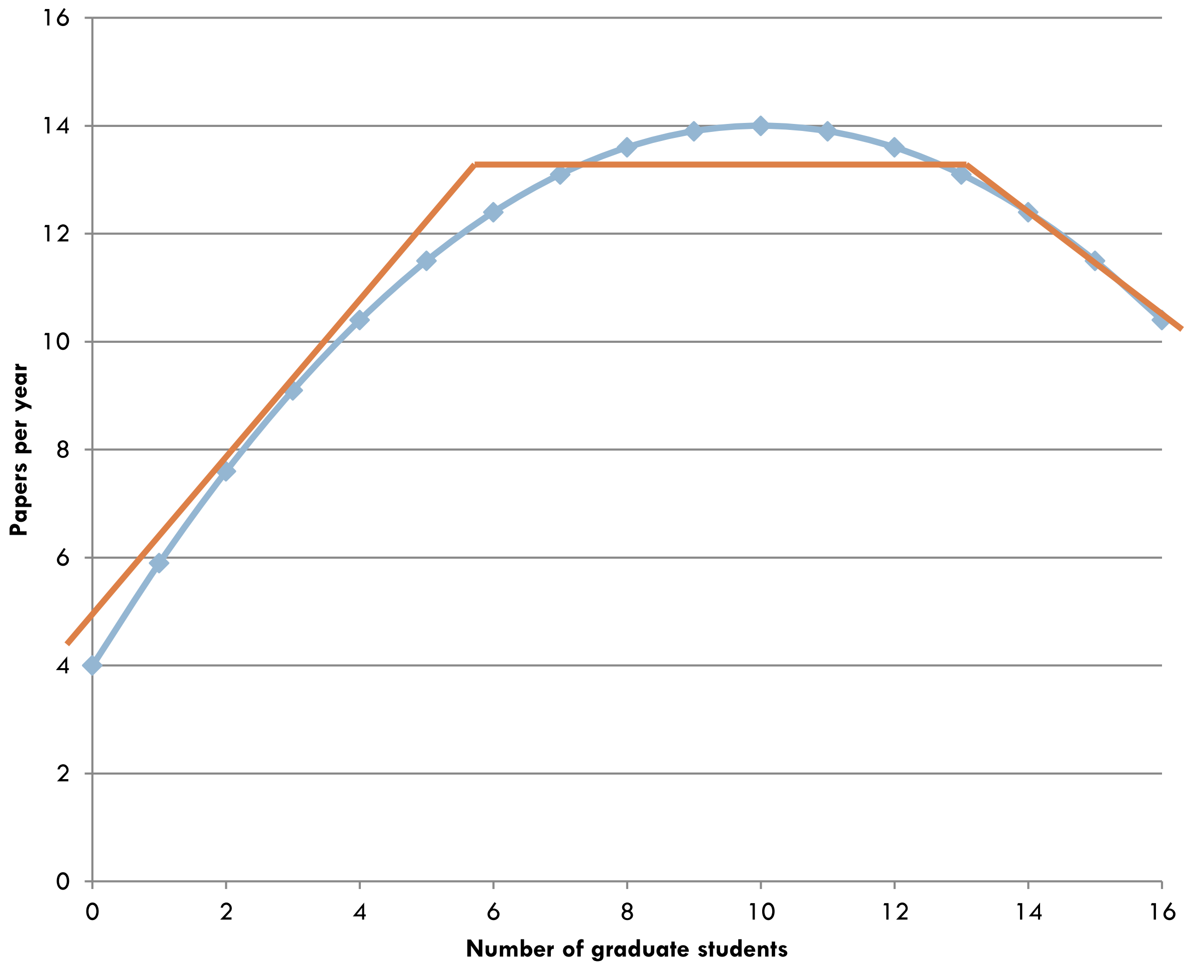
Regression Trees (non-linear; RepTree)

- If X>3
  - Y = 2
  - else If X<-7
    - Y = 4
    - Else Y = 3

Linear Regression Trees (linear; M5')

- If X>3
  - Y = 2A + 3B
  - else If X< -7
    - Y = 2A – 3B
    - Else Y = 2A + 0.5B + C

Linear Regression Tree



Later Lectures

- Other regressors
- Goodness metrics for comparing regressors
- Validating regressors

Next Lecture

- Classifiers – another type of prediction model