# CS 360: Machine Learning

Prof. Sara Mathieson

Fall 2020

HAVERFORD COLLEGE

# Admin

- Lab 2 due **THURSDAY night**
  - Grace period til Friday at 5
  - Continuous features optional
  - Information gain optional (can do cond. entropy instead)
- Office hours today **4:30-6pm**
- **Reading for Friday**
  - Duame 7.6 (2+ pages)
  - ISL 59-63 (4+ pages)

# Outline for September 22

- Finish Decision Trees (recap continuous features)
- Learning problem so far + terminology
- Bias-Variance tradeoff
- Linear regression

# Linear Regression Goals

- Regression as a way to study *expected loss* and the *bias-variance tradeoff*
- Review matrix algebra and expected values
- As an introduction to optimization (specifically *stochastic gradient descent*)

# Outline for September 22

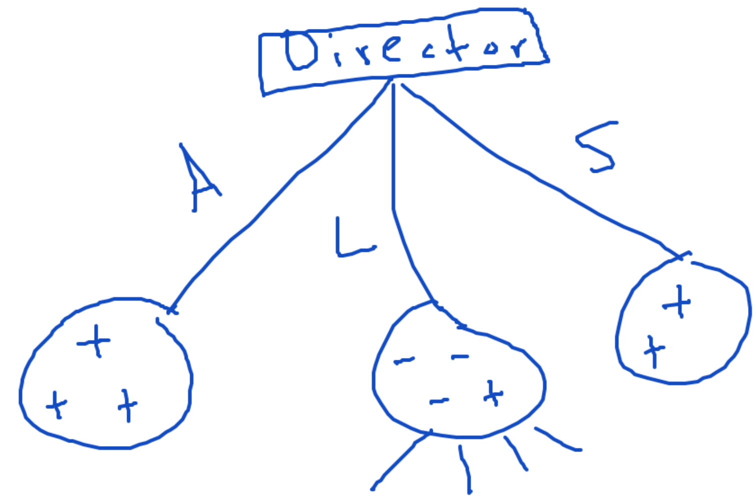
- Finish Decision Trees (recap continuous features)
- Learning problem so far + terminology
- Bias-Variance tradeoff
- Linear regression

# Handout 3

| Movie | Type | Length | Director | Famous actors | Liked? |
|---|---|---|---|---|---|
| m1 | Comedy | Short | Adamson | No | Yes |
| m2 | Animated | Short | Lasseter | No | No |
| m3 | Drama | Medium | Adamson | No | Yes |
| m4 | Animated | Long | Lasseter | Yes | No |
| m5 | Comedy | Long | Lasseter | Yes | No |
| m6 | Drama | Medium | Singer | Yes | Yes |
| m7 | Animated | Short | Singer | No | Yes |
| m8 | Comedy | Long | Adamson | Yes | Yes |
| m9 | Drama | Medium | Lasseter | No | Yes |

Y

root




$P(\text{Li} = \text{yes}) = \frac{2}{3}$

$H(\text{Li}) = -\left(\frac{2}{3}\log\frac{2}{3} + \frac{1}{3}\log\frac{1}{3}\right) = 0.92$

$H(\text{Li} \mid \text{T}) = 0.61$

$H(\text{Li} \mid \text{Le}) = 0.61$

$\boxed{H(\text{Li} \mid \text{D}) = 0.36}$ MIN ENTROPY

$H(\text{Li} \mid \text{F}) = 0.85$

$\text{Gain}(\text{Li}, \text{T}) = 0.92 - 0.61 = 0.31$

$\text{Gain}(\text{Li}, \text{Le}) = 0.92 - 0.61 = 0.31$

$\boxed{\text{Gain}(\text{Li}, \text{D}) = 0.92 - 0.36 = 0.56}$ MAX INFO GAIN

$\text{Gain}(\text{Li}, \text{F}) = 0.92 - 0.85 = 0.07$

# Handout 3

| Movie | Type | Length | Director | Famous actors | Liked? |
|---|---|---|---|---|---|
| m1 | Comedy | Short | Adamson | No | Yes |
| m2 | Animated | Short | Lasseter | No | No |
| m3 | Drama | Medium | Adamson | No | Yes |
| m4 | Animated | Long | Lasseter | Yes | No |
| m5 | Comedy | Long | Lasseter | Yes | No |
| m6 | Drama | Medium | Singer | Yes | Yes |
| m7 | Animated | Short | Singer | No | Yes |
| m8 | Comedy | Long | Adamson | Yes | Yes |
| m9 | Drama | Medium | Lasseter | No | Yes |

$$0.36$$

$$-\left(\frac{1}{4}\log\frac{1}{4}+\frac{3}{4}\log\frac{3}{4}\right)$$

weighted avg

$$H(Li \mid D) = \frac{3}{9} H(Li \mid D=A) + \frac{4}{9} \cdot H(Li \mid D=L) + \frac{2}{9} H(Li \mid D=S)$$

0    0

$$H(Li \mid D=A) = -P(Y \mid D=A)\log P(Y \mid D=A) - P(N \mid D=A)\underbrace{\log P(N \mid D=A)}_{\text{und}}$$

0    0

$$P(Y \mid D=A) = \frac{P(Y \text{ and } D=A)}{P(D=A)} = \frac{3}{3} = 1$$

# Handout 3

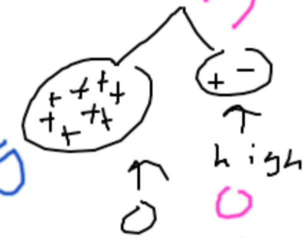
| Movie | Type | Length | Director | Famous actors | Liked? |
|---|---|---|---|---|---|
| m1 | Comedy | Short | Adamson | No | Yes |
| m2 | Animated | Short | Lasseter | No | No |
| m3 | Drama | Medium | Adamson | No | Yes |
| m4 | Animated | Long | Lasseter | Yes | No |
| m5 | Comedy | Long | Lasseter | Yes | No |
| m6 | Drama | Medium | Singer | Yes | Yes |
| m7 | Animated | Short | Singer | No | Yes |
| m8 | Comedy | Long | Adamson | Yes | Yes |
| m9 | Drama | Medium | Lasseter | No | Yes |

Y

0.36


$$-\left(\frac{1}{4}\log\frac{1}{4}+\frac{3}{4}\log\frac{3}{4}\right)$$

weighted avg




$$H(Li\mid D)=\frac{3}{9}H(Li\mid D=A)+\frac{4}{9}\cdot H(Li\mid D=L)+\frac{2}{9}H(Li\mid D=S)$$

0

Y output

$$H(Li\mid D=A)=-P(Yes\mid D=A)\log P(Yes\mid D=A)-P(No\mid D=A)\log P(No\mid D=A)$$

{Yes, No}

$\frac{0}{3}$ 0 und $\frac{0}{3}$

1 0

$$P(Y\mid D=A)=\frac{P(Y \text{ and } D=A)}{P(D=A)}=\frac{3}{3}=1$$

# Outline for September 22

- Finish Decision Trees (recap continuous features)

main

dtree = DecisionTree(train)

- Learning problem so far + terminology

"Sun"
"rain"

- Bias-Variance tradeoff

children
key: string
value: DTree

- Linear regression

newpart = Partition( )
child = DTree(newpart, d+1)
self.children[v] = child

# Continuous Features

(do this for the TRAIN only!)

| X | Y |
|---|---|
| 10 | Y |
| 7 | Y |
| 8 | N |
| 3 | Y |
| 7 | N |
| 12 | Y |
| 2 | Y |

1) Sort examples based on given feature

| 2 | 3 | 7 | 7 | 8 | 10 | 12 |
|---|---|---|---|---|---|---|
| Y | Y | Y | N | N | Y | Y |

2) Different label with same feature value, collapse to "None"

| 2 | 3 | 7 | 8 | 10 | 12 |
|---|---|---|---|---|---|
| Y | Y | None | N | Y | Y |

3) Whenever label changes, make a feature (use avg)

$\boxed{x \le 5}$ $\boxed{x \le 7.5}$ $\boxed{x \le 9}$

$x \le 5$

F
F
F

T
F
F
T

# Continuous Features (pair exercise)

(do this for the TRAIN only!)

3 new cols

| $x \le 44$ | temp | Y |
|---|---|---|
| F | 80 | Y |
| F | 48 | Y |
| F | 60 | N |
| F | 48 | Y |
| T | 40 | N |
| F | 48 | Y |
| F | 90 | Y |

1) Sort examples based on feature "temp"

| 40 | 48 | 48 | 48 | 60 | 80 | 90 |
|---|---|---|---|---|---|---|
| N | Y | Y | ~~N~~ Y | ~~Y~~ N | Y | ~~N~~ Y |

2) Different label with same feature value, collapse to "None"

| 40 | 48 | 60 | 80 | 90 |
|---|---|---|---|---|
| N | ~~None~~ Y | ~~Y~~ N | Y | ~~N~~ Y |

3) Whenever label changes, make a feature (use avg)

$x \le 44$  $x \le 54$  $x \le 70$

# Any other Lab 2 questions?

# Outline for September 22

- Finish Decision Trees (recap continuous features)
- Learning problem so far + terminology
- Bias-Variance tradeoff
- Linear regression

# Loss Functions

- E.g., zero-one loss
  - Simple accuracy - is prediction right?
  - For binary or multi-class prediction

$$l(y, \hat{y}) = \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & \text{otherwise} \end{cases}$$

- E.g., squared loss
  - For regression

$$l(y, \hat{y}) = (y - \hat{y})^2$$

- Absolute loss (also for regression)

$$\ell(y, \hat{y}) = |y - \hat{y}|$$

# Formalizing the learning problem

- Given:
  - Loss function, $\ell$
  - A sample of data $D$ from an unknown distribution of all data $\mathcal{D}$
  - A hypothesis space $H = \{h | h : X \to Y\}$
- Do:
  - Find a function $f(X) \to y$ that
  - minimize error over $\mathcal{D}$ with respect to $\ell$

# Why might learning fail?

# Inductive Bias

Training Data

class A

class B

Testing Data

Figures 2.1 and 2.2 from Duame

# Inductive Bias

Training Data

class A

class B

Testing Data

A

A

B

B

A: “fly”

B: “no fly”

Figures 2.1 and 2.2 from Duame

# Inductive Bias

Training Data

class A

class B

Testing Data


A: “bird”

B: “mammal”

Figures 2.1 and 2.2 from Duame

# Why might learning fail?

- Noise in the training data
  - Typos in a restaurant review
- Available features are insufficient
  - x-ray does not capture the medical issue
- “Correct” prediction is up to interpretation
  - Parental controls on web content
- Learning algorithm cannot cope with the data

# Hyperparameters

- Difficult to define precisely, but typically a parameter that controls other parameters

- What is one hyperparameter in decision trees?

Max depth!

- We can't choose hyperparameters via test data (breaks cardinal rule!)

- But we can use *validation data*

# General approach to training

1. Split your data into 70% training data, 10% development data and 20% test data. (validation data)

2. For each possible setting of your hyperparameters:

   (a) Train a model using that setting of hyperparameters on the training data.

   (b) Compute this model's error rate on the development data.

3. From the above collection of models, choose the one that achieved the lowest error rate on development data.

4. Evaluate that model on the test data to estimate future test performance.

# Outline for September 22

- Finish Decision Trees (recap continuous features)
- Learning problem so far + terminology
- Bias-Variance tradeoff

whiteboard

- Linear regression

# Regression setup and Expected Value

Regression setup

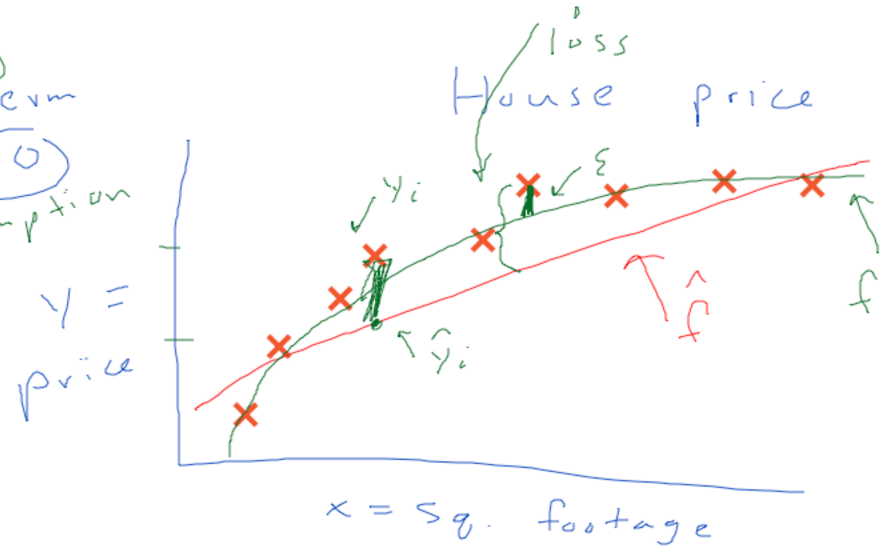
model: true $y = f(x) + \varepsilon$

error term (noise)

mean 0 assumption

$\hat{f}$ = estimate of $f$

$\hat{y} = \hat{f}(x)$ prediction




GOAL $\ell(y, \hat{y}) = (y - \hat{y})^2$

$E[(y - \hat{y})^2]$ expected loss

expected value = weighted avg

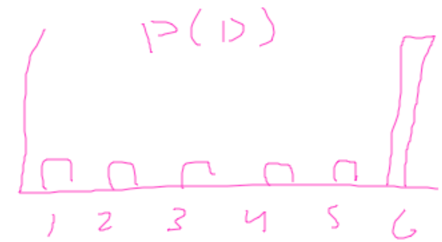
$$E[X] = \sum_{v} p(X = v) \cdot v$$

$$E(D) = \frac{1}{10}(1 + 2 \cdots 5) + \frac{1}{2} \cdot 6$$

$$= 4.5$$

$P(D = 6) = \frac{1}{2}$

$P(D \neq 6) = \frac{1}{10}$



# Compute Expected Loss

$$E[(y-\hat{y})^2] = E[(\underbrace{\overbrace{y - f}^{a}}_{\varepsilon} \underbrace{+}_{0} \overbrace{f - \hat{f}}^{b})^2]$$

$\hat{f}(x) = \hat{y}$

$$= \underbrace{\overbrace{Var(\varepsilon)}^{a^2}}_{\text{noise}} + \underbrace{\overbrace{E[(f-\hat{f})^2]}^{b^2}}_{\text{reducible error}}$$

noise

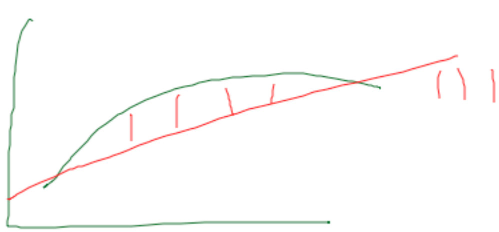
irreducible error

$(a+b)^2$

$= a^2 + b^2 + \overbrace{\cancel{2ab}}^{0}$

$Var(x) = E[(x-\mu)^2]$

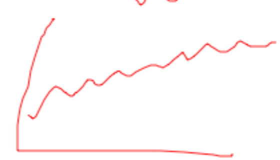
$$E[(f-\hat{f})^2] = E[(\underbrace{f - E[\hat{f}]}_{\text{bias}} \underbrace{+}_{0} \underbrace{E[\hat{f}] - \hat{f}}_{\text{variance}})^2]$$



bias

too simple

variance



too complex

# Outline for September 22

- Finish Decision Trees (recap continuous features)
- Learning problem so far + terminology
- Bias-Variance tradeoff
- Linear regression

# Assessing Model Accuracy

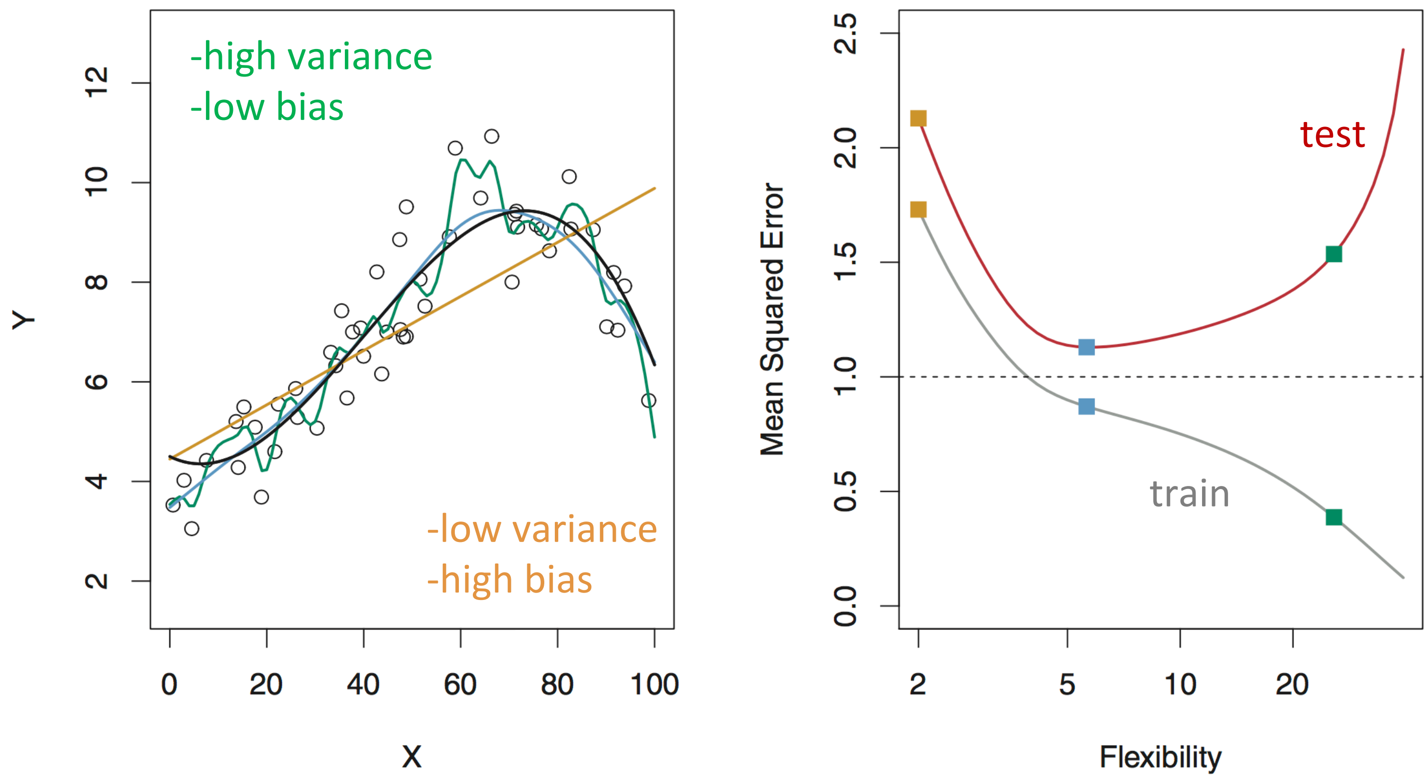



**FIGURE 2.9.** Left: *Data simulated from $f$, shown in black. Three estimates of $f$ are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.*

ISL, Fig 2.9

# Goals of Inference

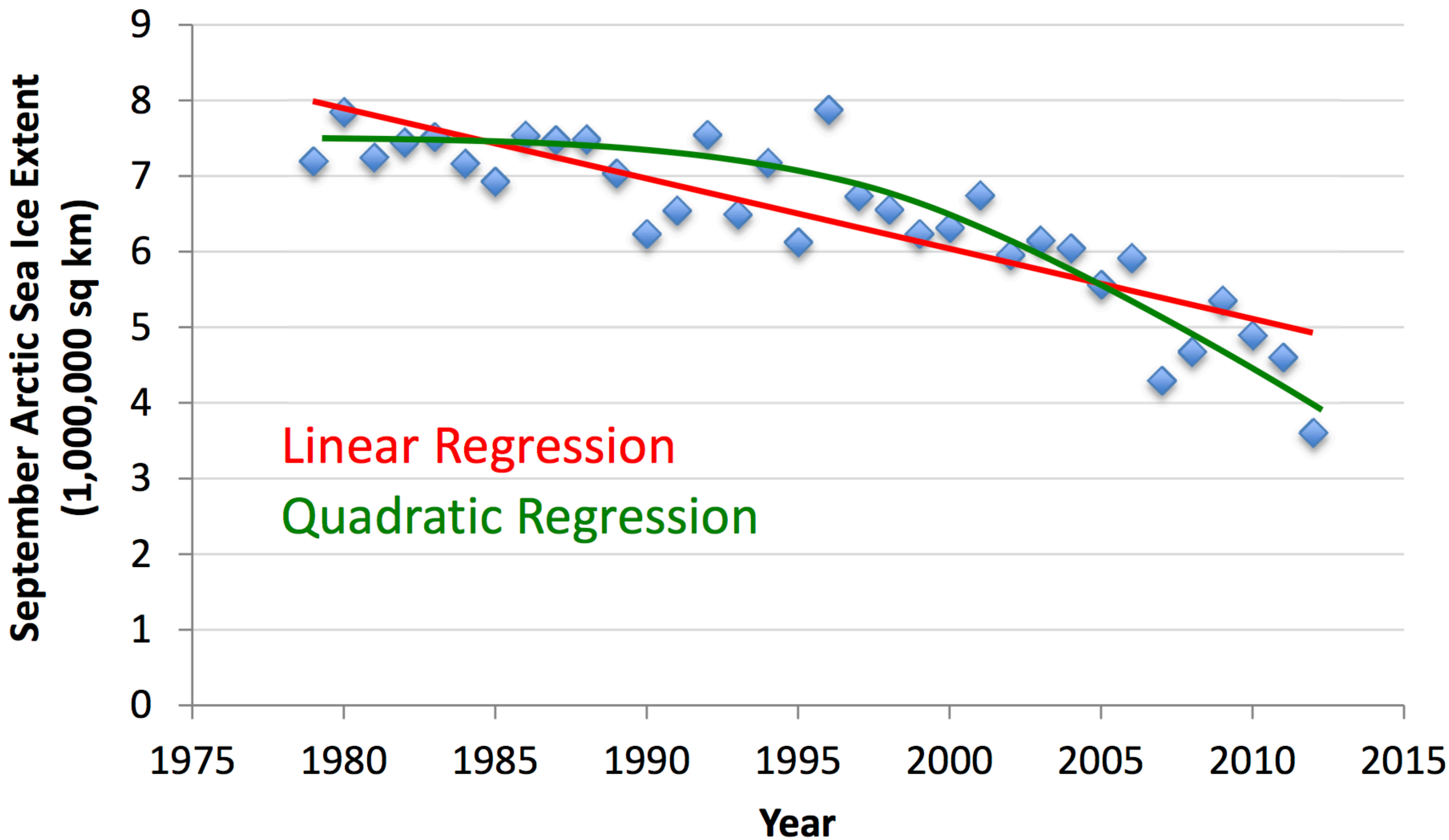
1) Which of the features/explanatory variables/predictors (x) are associated with the response variable (y)?

2) What is the relationship between x and y?

3) Is a linear model enough?

4) Can we predict y given a new x?

# Regression Example




Slide: modified from Jessica Wu

[Data from G. Witt. Journal of Statistics Education, Volume 21, Number 1 (2013)]