Predicting Real-valued outputs: an introduction to Regression

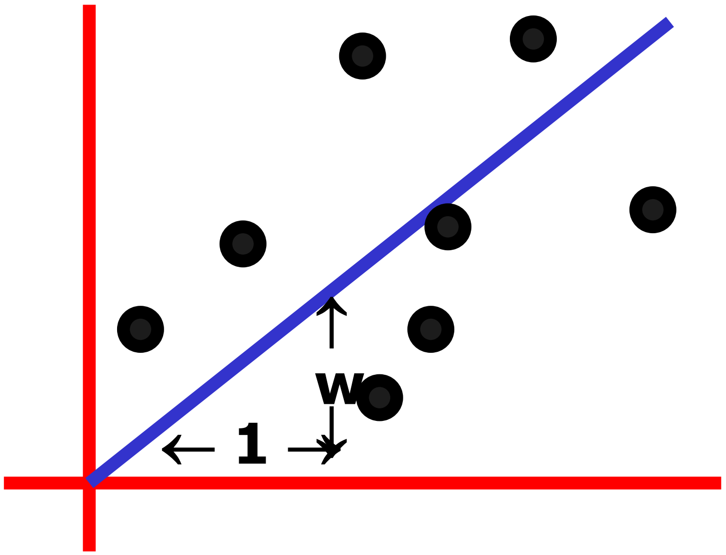
Note to other teachers and users of these slides. Andrew would be delighted if you found this source material useful in giving your own lectures. Feel free to use these slides verbatim, or to modify them to fit your own needs. PowerPoint originals are available. If you make use of a significant portion of these slides in your own lecture, please include this message, or the following link to the source repository of Andrew's tutorials: http://www.cs.cmu.edu/~awm/tutorials . Comments and corrections gratefully received.

Andrew W. Moore
Professor
School of Computer Science
Carnegie Mellon University
www.cs.cmu.edu/~awm
awm@cs.cmu.edu
412-268-7599

This is reordered material from the Neural Nets lecture and the "Favorite Regression Algorithms" lecture

Single-Parameter Linear Regression

Linear Regression



DATASET

inputs	outputs
$x_1 = 1$	$y_1 = 1$
$x_2 = 3$	$y_2 = 2.2$
$x_3 = 2$	$y_3 = 2$
$x_4 = 1.5$	$y_4 = 1.9$
$x_5 = 4$	$y_5 = 3.1$

Linear regression assumes that the expected value of the output given an input, *E[y|x]*, is linear.

Simplest case: Out(x) = wx for some unknown w.

Given the data, we can estimate w.

1-parameter linear regression

Assume that the data is formed by

$$y_i = wx_i + \text{noise}_i$$

where…

- the noise signals are independent
- the noise has a normal distribution with mean 0 and unknown variance σ^2

p(y|w,x) has a normal distribution with

- mean wx
- variance σ^2

Bayesian Linear Regression

p(y|w,x) = Normal (mean wx, var σ^2)

We have a set of datapoints (x_1,y_1) (x_2,y_2) … (x_n,y_n) which are EVIDENCE about w.

We want to infer w from the data.

$$p(w|x_1, x_2, x_3, \ldots x_n, y_1, y_2 \ldots y_n)$$

- You can use BAYES rule to work out a posterior distribution for w given the data.
- Or you could do Maximum Likelihood Estimation

Maximum likelihood estimation of *w*

Asks the question:

"For which value of *w* is this data most likely to have happened?"

<=>

For what *w* is

$p(y_1, y_2 \ldots y_n \mid x_1, x_2, x_3, \ldots x_n, w)$ maximized?

<=>

For what *w* is

$$\prod_{i=1}^{n} p(y_i \mid w, x_i) \text{ maximized}$$

For what *w* is

$$\prod_{i=1}^{n} p(y_i | w, x_i) \text{ maximized}$$

For what *w* is

$$\prod_{i=1}^{n} \exp\left(-\frac{1}{2}\left(\frac{y_i - wx_i}{\sigma}\right)^2\right) \text{ maximized?}$$

For what *w* is

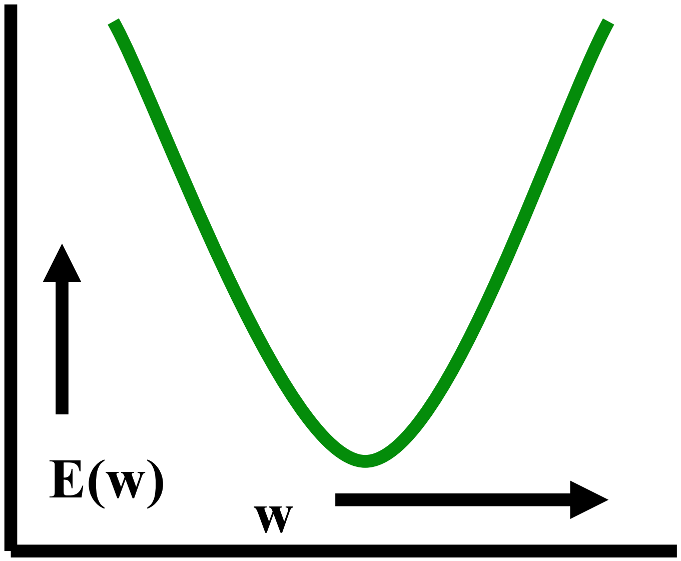
$$\sum_{i=1}^{n} -\frac{1}{2}\left(\frac{y_i - wx_i}{\sigma}\right)^2 \text{ maximized?}$$

For what *w* is

$$\sum_{i=1}^{n} \left(y_i - wx_i\right)^2 \text{ minimized?}$$

Linear Regression

The maximum likelihood *w* is the one that minimizes sum-of-squares of <u>residuals</u>



$$\mathrm{E} = \sum_i \left(y_i - w x_i\right)^2$$

$$= \sum_i y_i^2 - \left(2\sum x_i y_i\right)w + \left(\sum x_i^2\right)w^2$$

We want to minimize a quadratic function of *w*.

Linear Regression

Easy to show the sum of squares is minimized when

$$w = \frac{\sum x_i y_i}{\sum x_i^{\;2}}$$

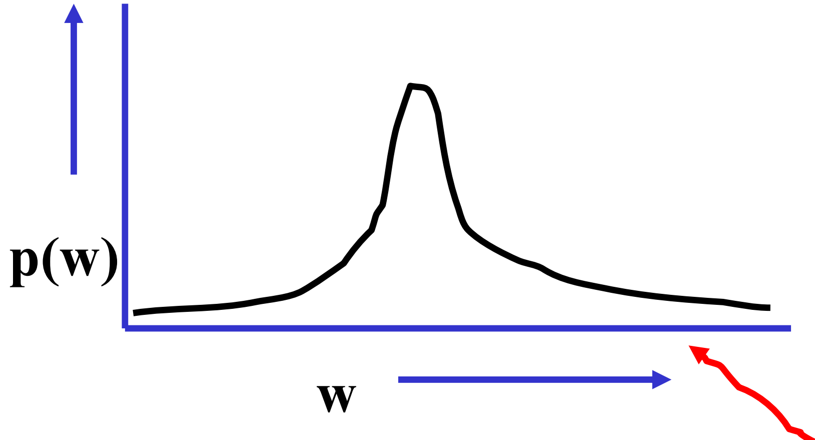
The maximum likelihood model is $\text{Out}(x) = wx$

We can use it for prediction

Linear Regression

Easy to show the sum of squares is minimized when

$$w = \frac{\sum x_i y_i}{\sum x_i^2}$$



Note: In Bayesian stats you'd have ended up with a prob dist of w

And predictions would have given a prob dist of expected output

Often useful to know your confidence. Max likelihood can give some kinds of confidence too.

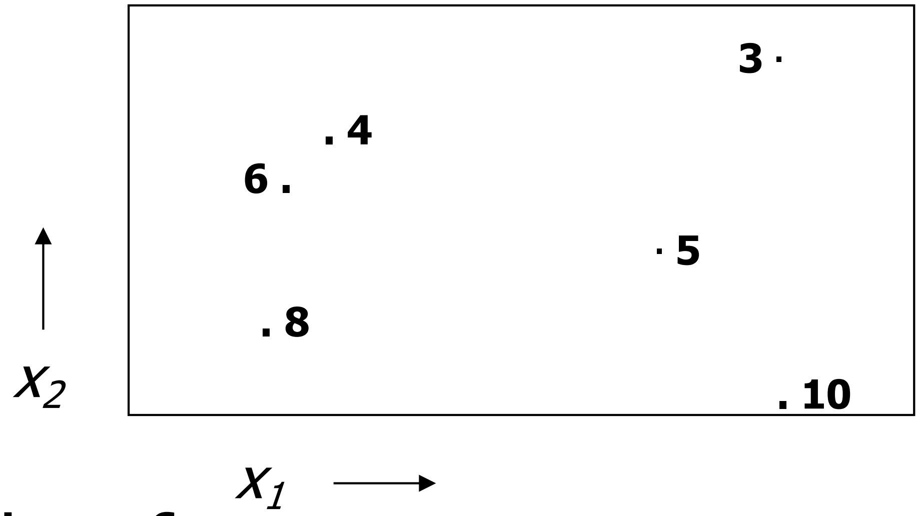
The maximum likelihood model is $\mathrm{Out}(x) = wx$

We can use it for prediction

Multivariate Linear Regression

Multivariate Regression

What if the inputs are vectors?



2-d input example

Dataset has form

$\mathbf{x}_1$	y_1
$\mathbf{x}_2$	y_2
$\mathbf{x}_3$	y_3
.:	:
.	
$\mathbf{x}_R$	y_R

Multivariate Regression

Write matrix X and Y thus:

$$\mathbf{x} = \begin{bmatrix} \ldots\mathbf{x}_1\ldots\ldots \\ \ldots\mathbf{x}_2\ldots\ldots \\ \mathrm{M} \\ \ldots\mathbf{x}_R\ldots\ldots \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1m} \\ x_{21} & x_{22} & \ldots & x_{2m} \\ & & \mathrm{M} & \\ x_{R1} & x_{R2} & \ldots & x_{Rm} \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \mathrm{M} \\ y_R \end{bmatrix}$$

(there are *R* datapoints. Each input has *m* components)

The linear regression model assumes a vector ***w*** such that

$$\mathrm{Out}(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} = w_1 x[1] + w_2 x[2] + \ldots . w_m x[D]$$

The max. likelihood ***w*** is $\boldsymbol{w} = (X^T X)^{-1}(X^T Y)$

Multivariate Regression

Write matrix X and Y thus:

$$\mathbf{x} = \begin{bmatrix} \ldots\mathbf{x}_1\ldots\ldots \\ \ldots\mathbf{x}_2\ldots\ldots \\ \mathrm{M} \\ \ldots\mathbf{x}_R\ldots\ldots \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1m} \\ x_{21} & x_{22} & \ldots & x_{2m} \\ & & \mathrm{M} & \\ x_{R1} & x_{R2} & \ldots & x_{Rm} \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \mathrm{M} \\ y_R \end{bmatrix}$$

(there are *R* datapoints. Each input

IMPORTANT EXERCISE: PROVE IT !!!!!

The linear regression model assumes a vector ***w*** such that

$$\mathrm{Out}(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} = w_1x[1] + w_2x[2] + \ldots.w_mx[D]$$

The max. likelihood ***w*** is $\boldsymbol{w} = (X^TX)^{-1}(X^TY)$

Multivariate Regression (con't)

The max. likelihood ***w*** is ***w*** = $(X^T X)^{-1}(X^T Y)$

$X^T X$ is an *m* x *m* matrix: i,j'th elt is $\sum_{k=1}^{R} x_{ki} x_{kj}$

$X^T Y$ is an *m*-element vector: i'th elt $\sum_{k=1}^{R} x_{ki} y_k$

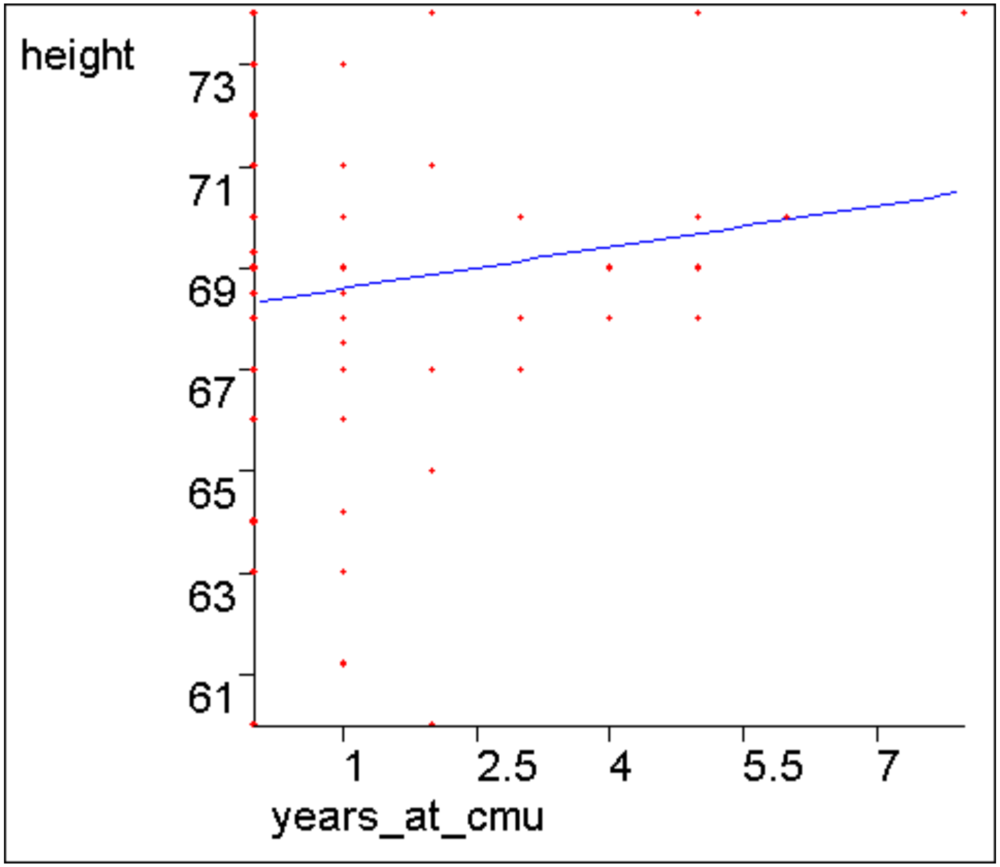
Constant Term in Linear Regression

What about a constant term?

We may expect linear data that does not go through the origin.

Statisticians and Neural Net Folks all agree on a simple obvious hack.

Can you guess??



The constant term

- The trick is to create a fake input "X_0" that always takes the value 1

X_1	X_2	Y
2	4	16
3	4	17
5	5	20

Before:

$Y=w_1X_1+ w_2X_2$

…has to be a poor model

X_0	X_1	X_2	Y
1	2	4	16
1	3	4	17
1	5	5	20

After:

$Y= w_0X_0+w_1X_1+ w_2X_2$

$= w_0+w_1X_1+ w_2X_2$

…has a fine constant term

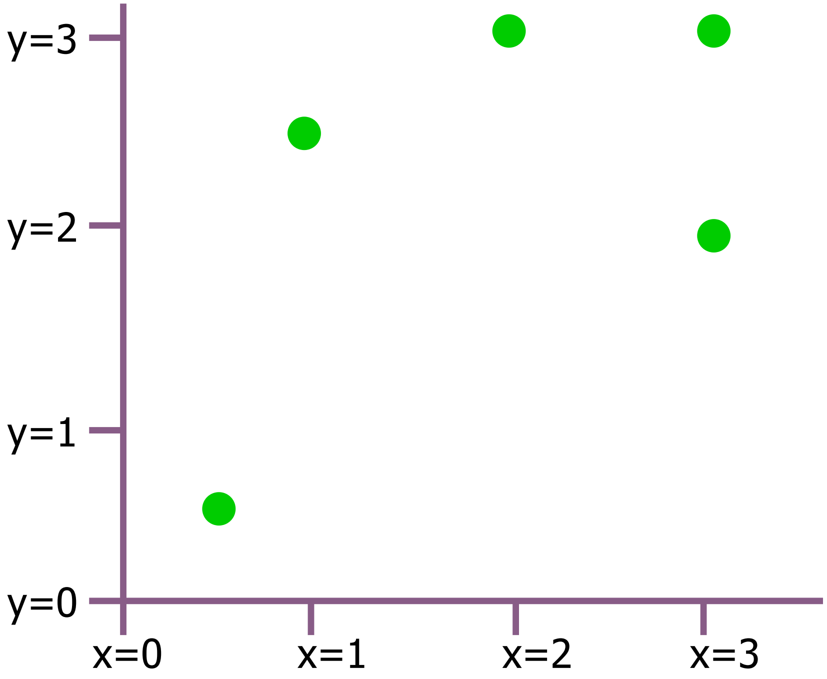
In this example, You should be able to see the MLE w_0 , w_1 and w_2 by inspection

Non-linear Regression

Non-linear Regression

- Suppose you know that y is related to a function of x in such a way that the predicted values have a non-linear dependence on w, e.g:

x_i	y_i
½	½
1	2.5
2	3
3	2
3	3



Assume $y_i \sim N(\sqrt{w + x_i}, \sigma^2)$

What's the MLE estimate of w?

Non-linear MLE estimation

$$\underset{w}{\operatorname{argmax}} \log p(y_1, y_2, \ldots, y_R \mid x_1, x_2, \ldots, x_R, \sigma, w) =$$

$$\underset{w}{\operatorname{argmin}} \sum_{i=1}^{R} \left(y_i - \sqrt{w + x_i} \right)^2 =$$

Assuming i.i.d. and then plugging in equation for Gaussian and simplifying.

$$\left(w \text{ such that } \sum_{i=1}^{R} \frac{y_i - \sqrt{w + x_i}}{\sqrt{w + x_i}} = 0 \right) =$$

Setting dLL/dw equal to zero

Non-linear MLE estimation

$$\underset{w}{\operatorname{argmax}} \log p(y_1, y_2, \ldots, y_R \mid x_1, x_2, \ldots, x_R, \sigma, w) =$$

$$\underset{w}{\operatorname{argmin}} \sum_{i=1}^{R} \left(y_i - \sqrt{w + x_i} \right)^2 =$$

Assuming i.i.d. and then plugging in equation for Gaussian and simplifying.

$$\left(w \text{ such that } \sum_{i=1}^{R} \frac{y_i - \sqrt{w + x_i}}{\sqrt{w + x_i}} = 0 \right) =$$

Setting dLL/dw equal to zero

We're down the algebraic toilet

So guess what we do?

Non-linear MLE estimation

$$\underset{w}{\operatorname{argmax}} \log p(y_1, y_2, \ldots, y_R \mid x_1, x_2, \ldots, x_R, \sigma, w) =$$

$$w + x_i \Big)^2 =$$

$$+ x_i \quad = 0 \Big) =$$

Assuming i.i.d. and then plugging in equation for Gaussian and simplifying.

Setting dLL/dw equal to zero

Common (but not only) approach:

Numerical Solutions:

- Line Search
- Simulated Annealing
- Gradient Descent
- Conjugate Gradient
- Levenberg Marquart
- Newton's Method

We're down the algebraic toilet

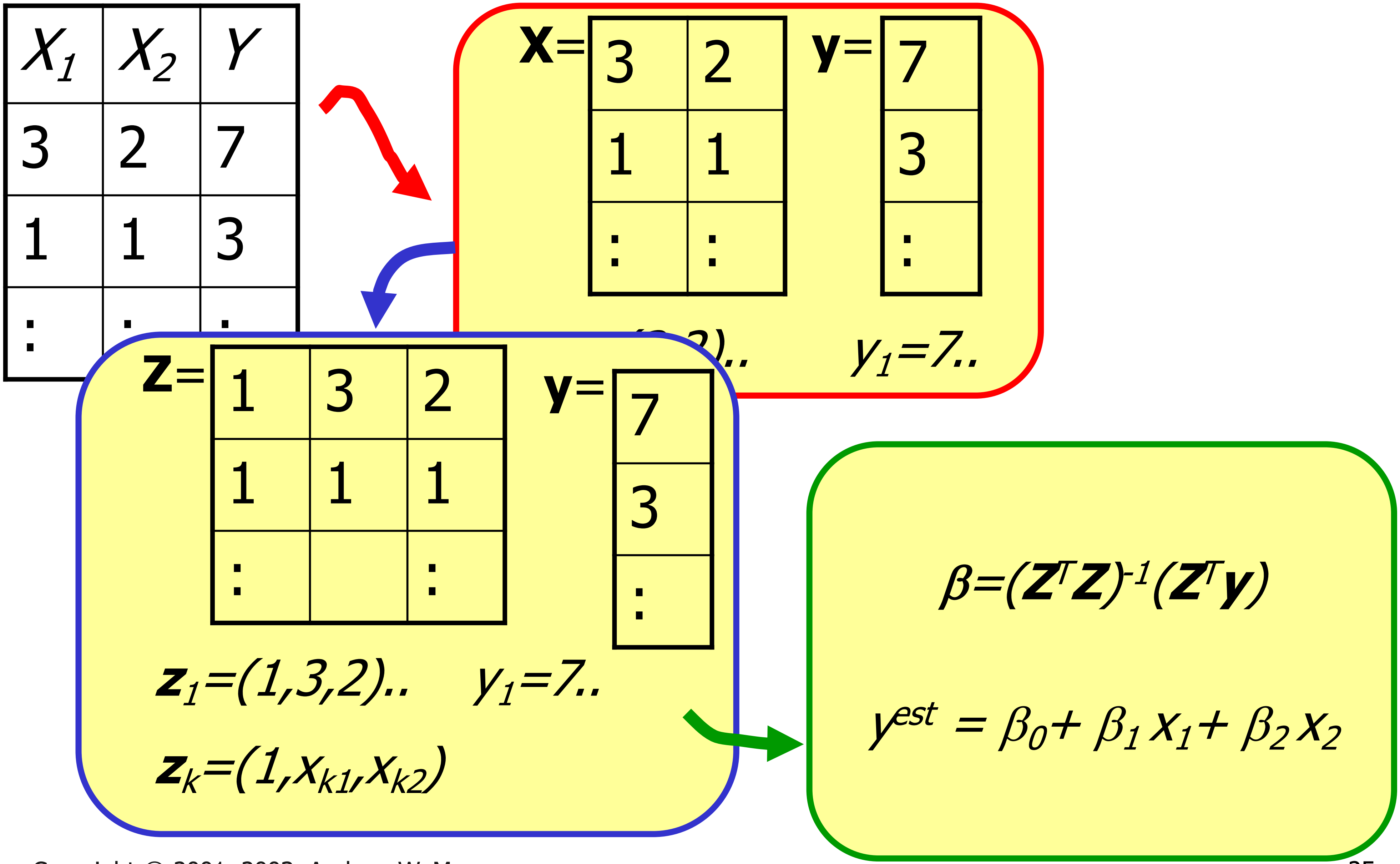
Also, special purpose statistical-optimization-specific tricks such as E.M. (See Gaussian Mixtures lecture for introduction)

So guess what we do?

Polynomial Regression

Polynomial Regression

So far we've mainly been dealing with linear regression



X_1	X_2	Y
3	2	7
1	1	3
:	:	:

$\mathbf{X}=$

3	2
1	1
:	:

$\mathbf{y}=$

7
3
:

(3,2).. $\quad y_1=7..$

$\mathbf{Z}=$

1	3	2
1	1	1
:		:

$\mathbf{y}=$

7
3
:

$\mathbf{z}_1=(1,3,2).. \quad y_1=7..$

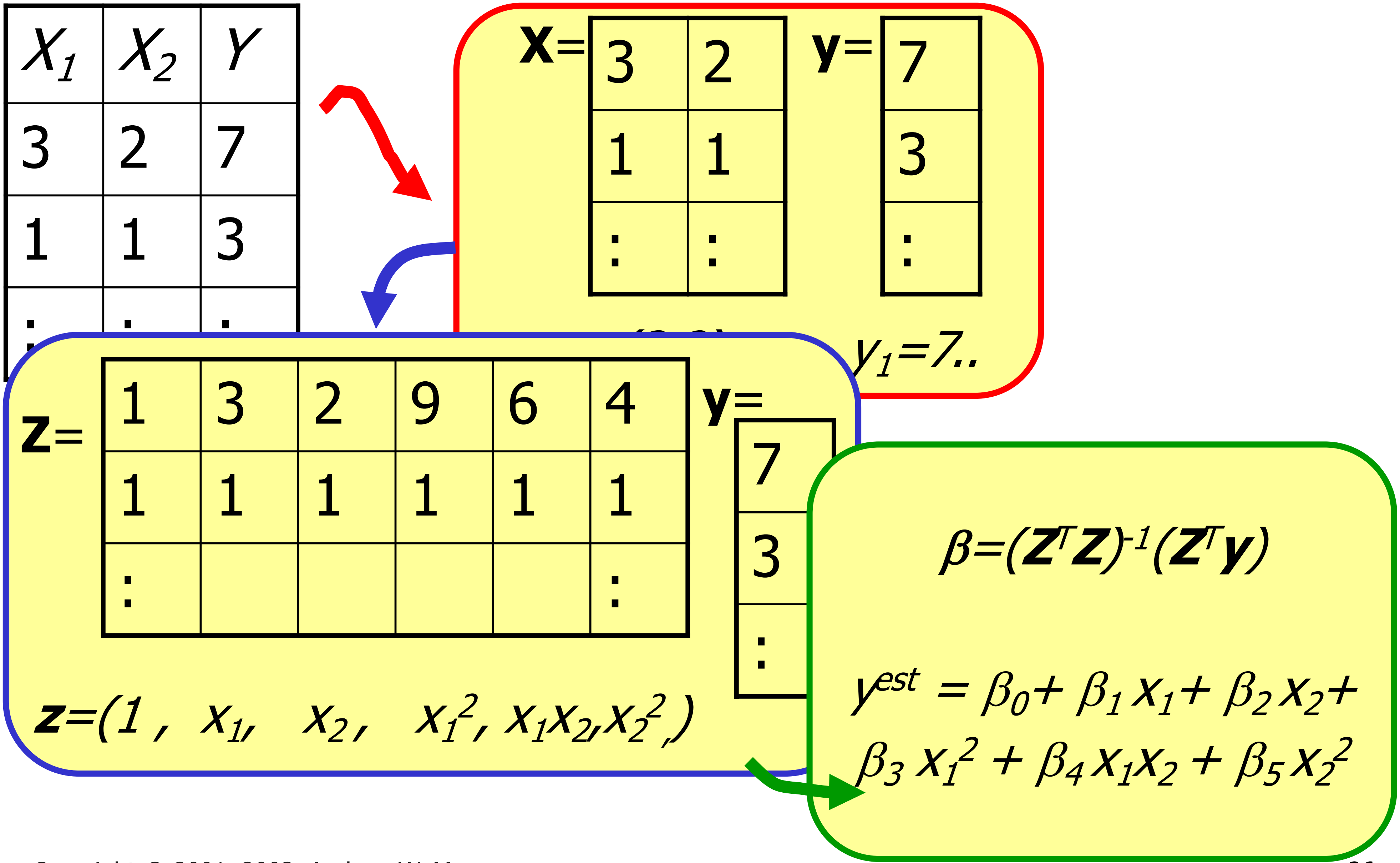
$\mathbf{z}_k=(1,x_{k1},x_{k2})$

$$\beta=(\mathbf{Z}^T\mathbf{Z})^{-1}(\mathbf{Z}^T\mathbf{y})$$

$$y^{est} = \beta_0+ \beta_1 x_1+ \beta_2 x_2$$

Quadratic Regression

It's trivial to do linear fits of fixed nonlinear basis functions



X_1	X_2	Y
3	2	7
1	1	3
:	:	:

$\mathbf{X}=$

3	2
1	1
:	:

$\mathbf{y}=$

7
3
:

$y_1=7$..

$\mathbf{Z}=$

1	3	2	9	6	4
1	1	1	1	1	1
:					:

$\mathbf{y}=$

7
3
:

$\mathbf{z}=(1\,,\ x_1,\ \ x_2\,,\ \ x_1^2, x_1x_2,x_2^2,)$

$$\beta=(\mathbf{Z}^T\mathbf{Z})^{-1}(\mathbf{Z}^T\mathbf{y})$$

$$y^{est} = \beta_0+ \beta_1 x_1+ \beta_2 x_2+ \beta_3 x_1^2 + \beta_4 x_1x_2 + \beta_5 x_2^2$$

Quadratic Regression

It's triv

X_1	X_2
3	2
1	1
:	:

$\mathbf{z}$=

1
1
:

$\mathbf{z}=(1$,

Each component of a z vector is called a term.

Each column of the Z matrix is called a term column

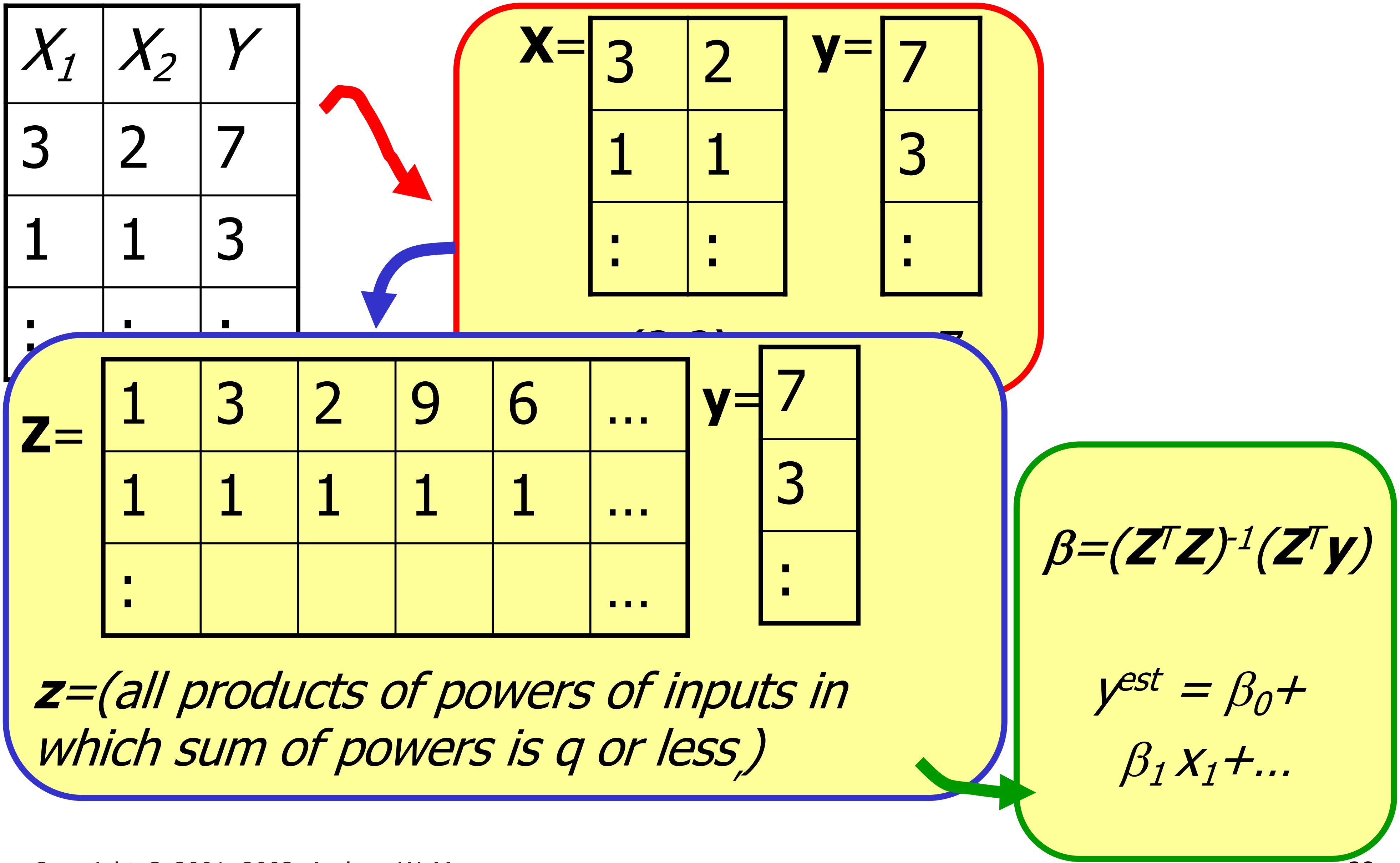
How many terms in a quadratic regression with *m* inputs?

- 1 constant term
- m linear terms
- (m+1)-choose-2 = m(m+1)/2 quadratic terms

(m+2)-choose-2 terms in total $= O(m^2)$

Note that solving $\beta=(\mathbf{Z}^T\mathbf{Z})^{-1}(\mathbf{Z}^T\mathbf{y})$ is thus $O(m^6)$

Q^{th}-degree polynomial Regression



X_1	X_2	Y
3	2	7
1	1	3
:	:	:

$$\mathbf{X}=\begin{bmatrix}3 & 2\\ 1 & 1\\ : & :\end{bmatrix} \quad \mathbf{y}=\begin{bmatrix}7\\ 3\\ :\end{bmatrix}$$

$$\mathbf{Z}=\begin{bmatrix}1 & 3 & 2 & 9 & 6 & \ldots\\ 1 & 1 & 1 & 1 & 1 & \ldots\\ : & & & & & \ldots\end{bmatrix} \quad \mathbf{y}=\begin{bmatrix}7\\ 3\\ :\end{bmatrix}$$

***z**=(all products of powers of inputs in which sum of powers is q or less,)*

$$\beta=(\mathbf{Z}^T\mathbf{Z})^{-1}(\mathbf{Z}^T\mathbf{y})$$

$$y^{est} = \beta_0+ \beta_1 x_1+\ldots$$

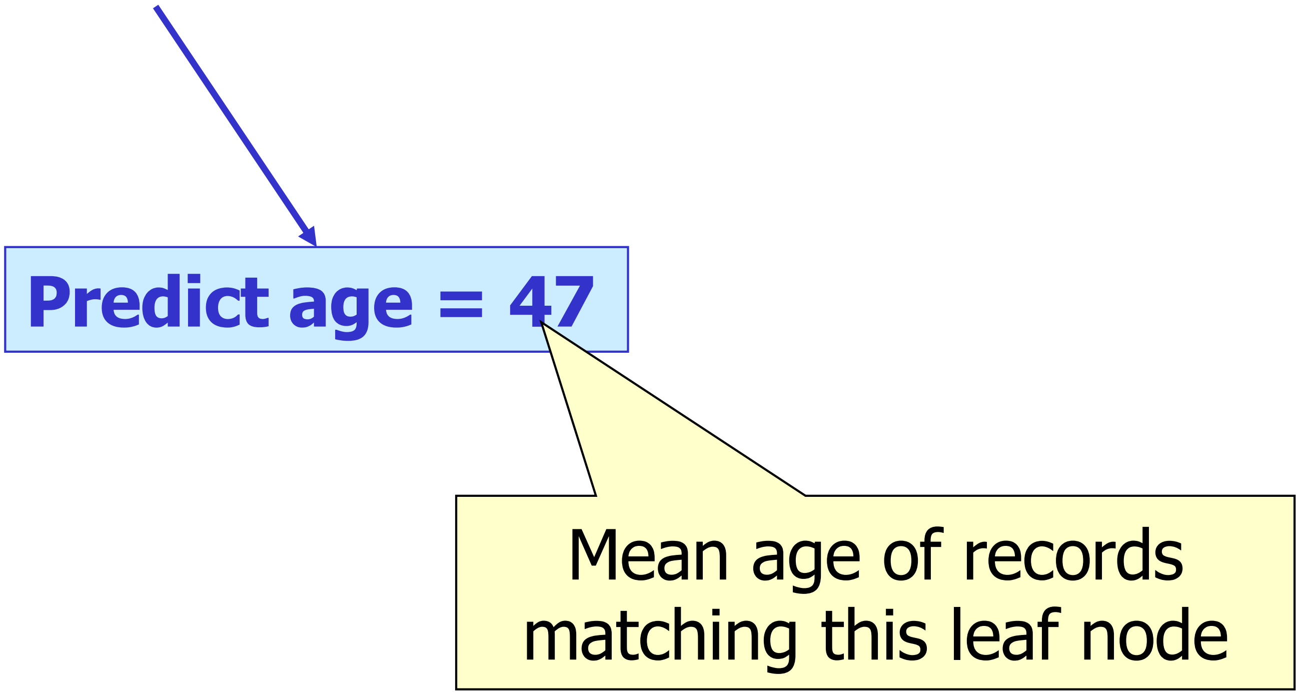
Copyright © 2001, 2003, Andrew W. Moore

Regression Trees

Regression Trees

- "Decision trees for regression"

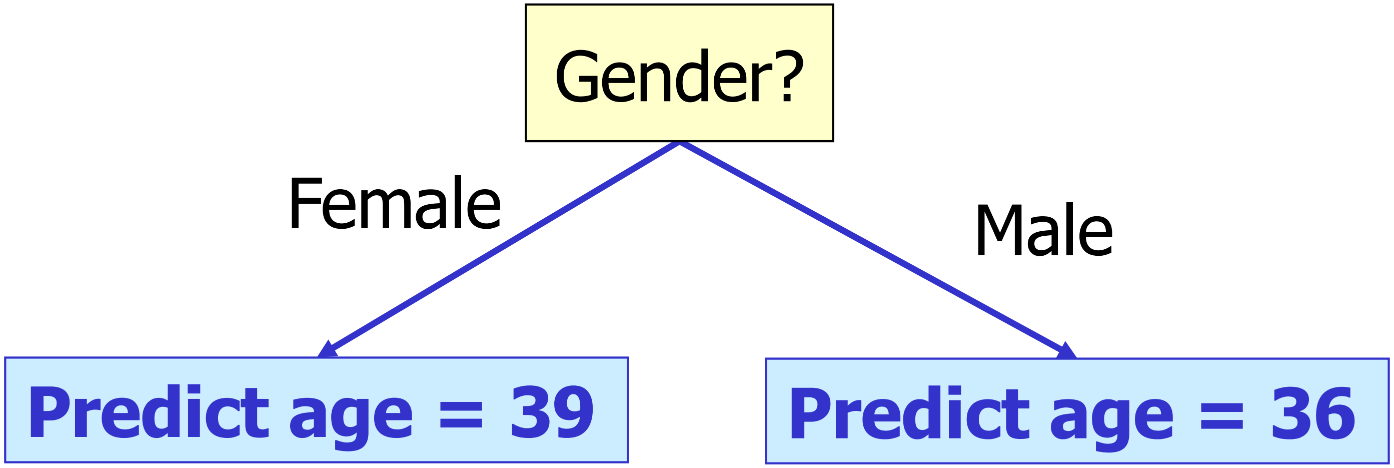
A regression tree leaf



Predict age = 47

Mean age of records matching this leaf node

A one-split regression tree



Gender?

Female

Male

Predict age = 39

Predict age = 36

Choosing the attribute to split on

Gender	Rich?	Num. Children	Num. Beany Babies	Age
Female	No	2	1	38
Male	No	0	0	24
Male	Yes	0	5+	72
:	:	:	:	:

- We can’t use information gain.
- What should we use?

Choosing the attribute to split on

Gender	Rich?	Num. Children	Num. Beany Babies	Age
Female	No	2	1	38
Male	No	0	0	24
Male	Yes	0	5+	72
:	:	:	:	:

MSE(Y|X) = The expected squared error if we must predict a record's Y value given only knowledge of the record's X value

If we're told *x=j*, the smallest expected error comes from predicting the mean of the Y-values among those records in which *x=j*. Call this mean quantity $\mu_y^{x=j}$

Then…

$$MSE(Y|X) = \frac{1}{R}\sum_{j=1}^{N_X}\sum_{(k \text{ such that } x_k = j)}(y_k - \mu_y^{x=j})^2$$

Choosing the attribute to split on

Gender	Rich?	Num. Children	Num. Beany Babies	Age
Female	N			
Male	N			
Male	Y			
:	:			

Regression tree attribute selection: greedily choose the attribute that minimizes MSE(Y|X)

Guess what we do about real-valued inputs?

Guess how we prevent overfitting

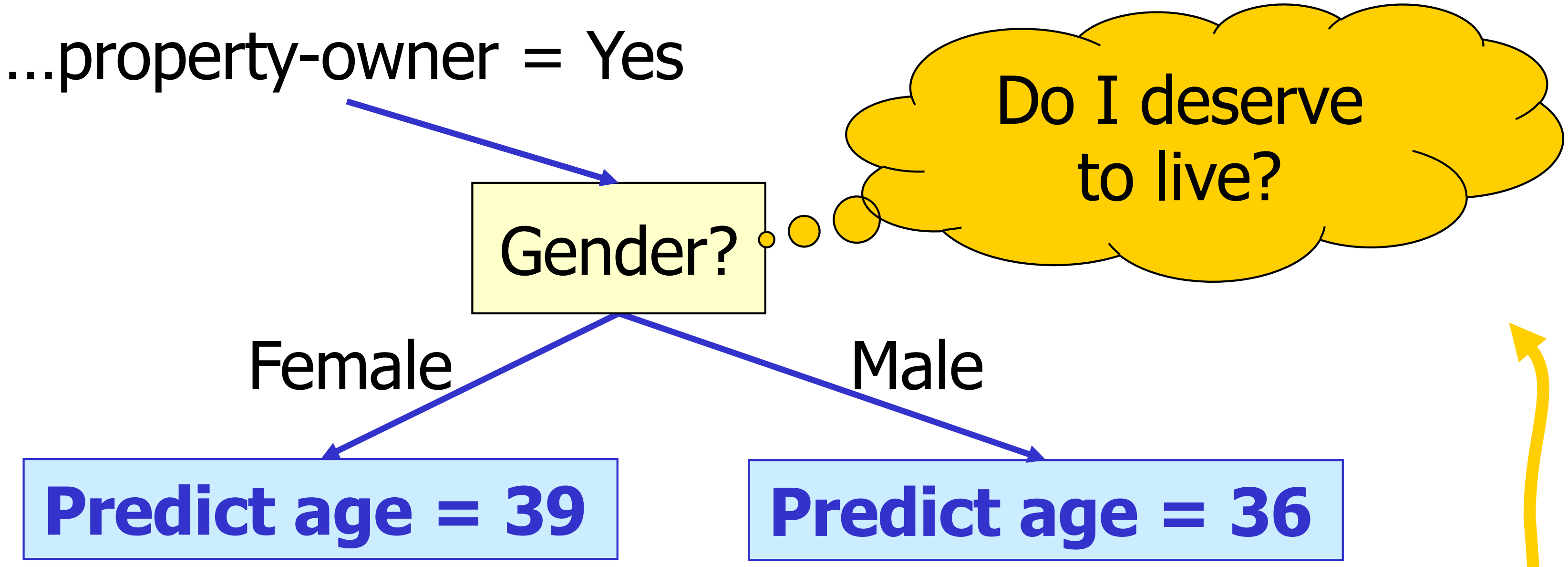
MSE(Y|X) = The expected squared error if we must predict a record's Y value given only knowledge of the record's X value

If we're told *x=j*, the smallest expected error comes from predicting the mean of the Y-values among those records in which *x=j*. Call this mean quantity $\mu_y^{x=j}$

Then…

$$MSE(Y|X) = \frac{1}{R}\sum_{j=1}^{N_X} \sum_{(k \text{ such that } x_k = j)} (y_k - \mu_y^{x=j})^2$$

Pruning Decision



…property-owner = Yes

Gender?

Do I deserve to live?

Female

Male

Predict age = 39

Predict age = 36

# property-owning females = 56712

Mean age among POFs = 39

Age std dev among POFs = 12

# property-owning males = 55800

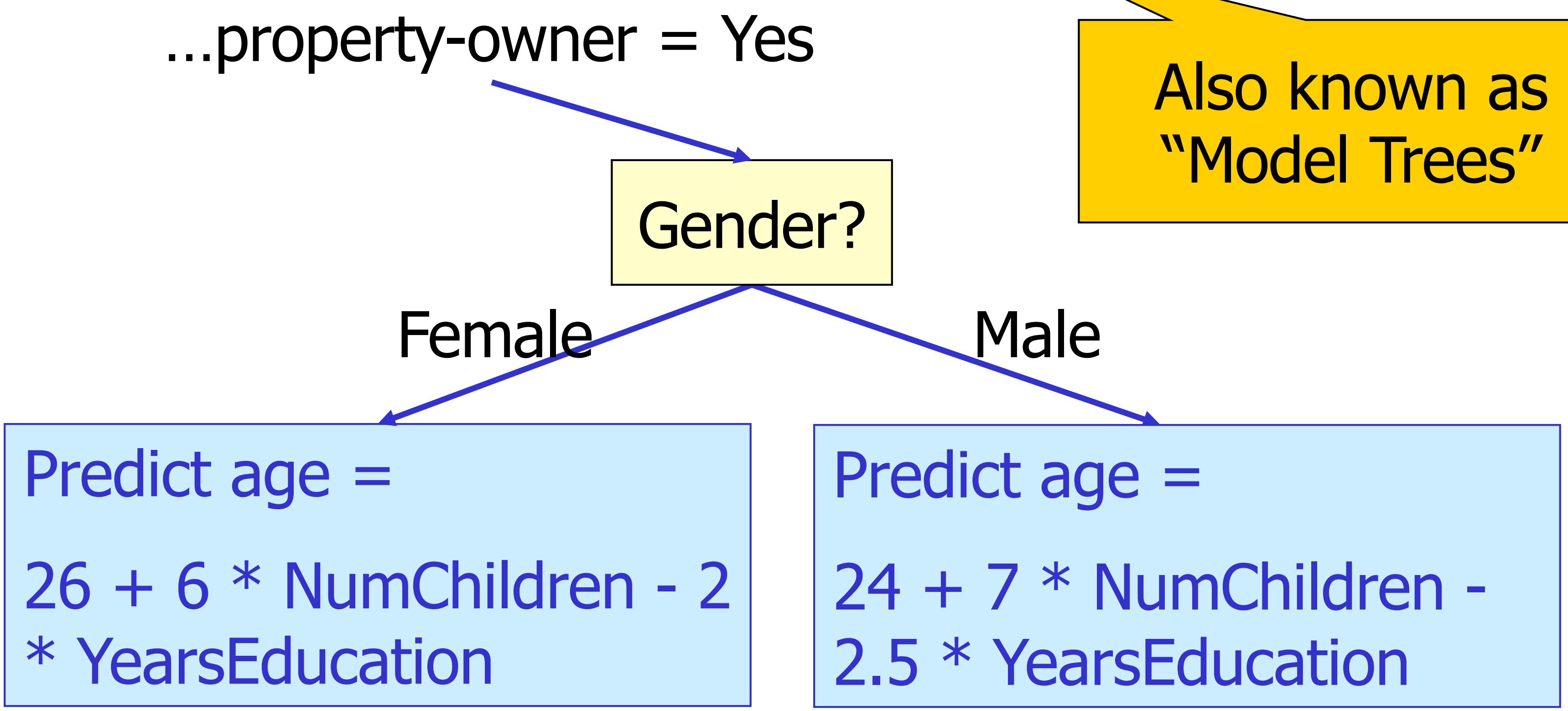
Mean age among POMs = 36

Age std dev among POMs = 11.5

Use a standard Chi-squared test of the null-hypothesis "these two populations have the same mean" and Bob's your uncle.

Linear Regression Trees

Also known as "Model Trees"



…property-owner = Yes

Gender?

Female

Male

Predict age =

26 + 6 * NumChildren - 2 * YearsEducation

Predict age =

24 + 7 * NumChildren - 2.5 * YearsEducation

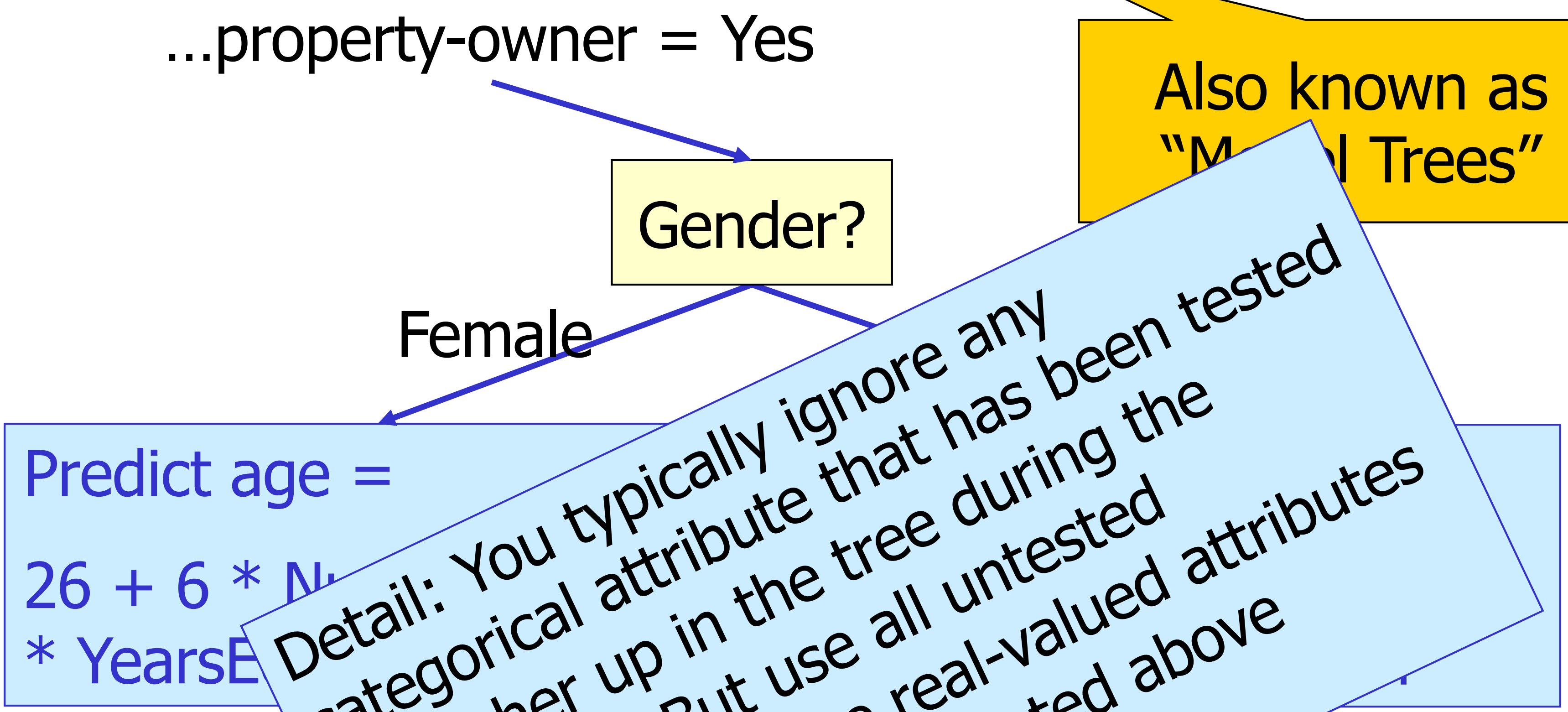
Leaves contain linear functions (trained using linear regression on all records matching that leaf)

Split attribute chosen to minimize MSE of regressed children.

Pruning with a different Chi-squared

Linear Regression Trees

Also known as "M...l Trees"



Detail: You typically ignore any categorical attribute that has been tested on higher up in the tree during the regression. But use all untested attributes, and use real-valued attributes even if they've been tested above

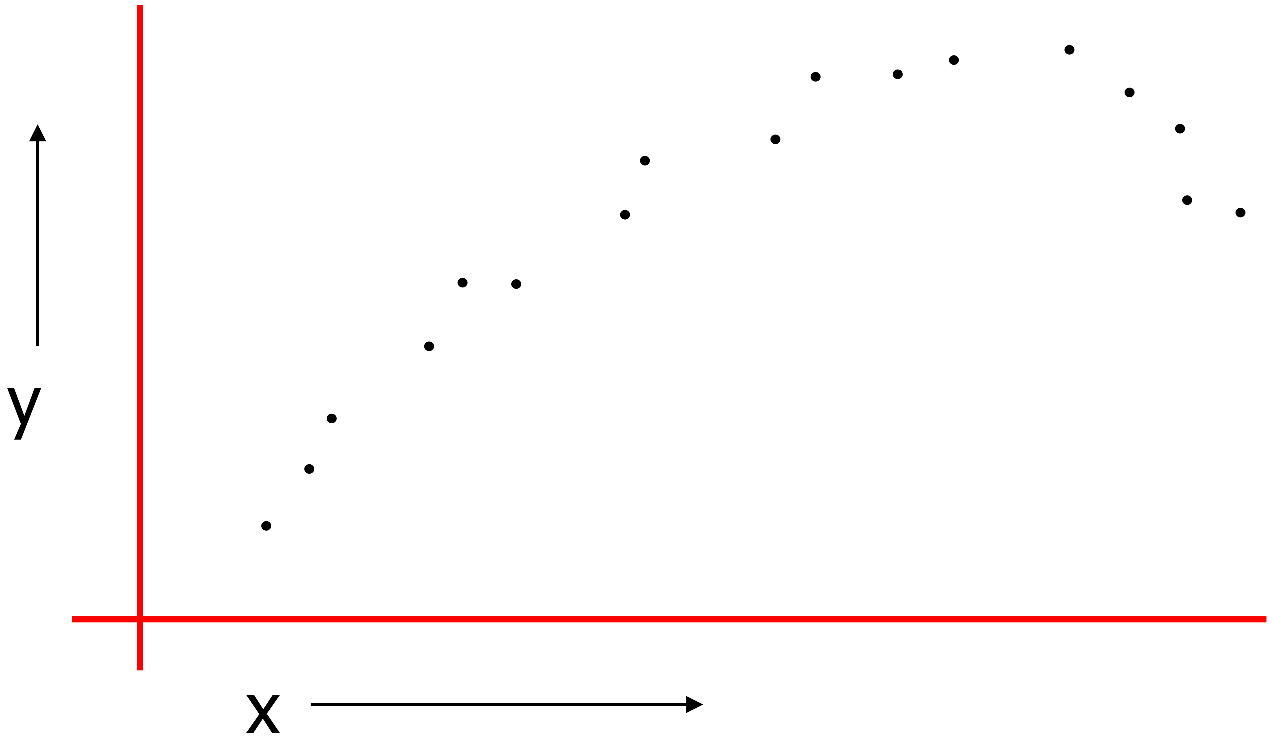
Leaves contain … functions (traine… linear regression … records matching t…)

…e chosen to minimize …regressed children.

Pruning with a different Chi-squared

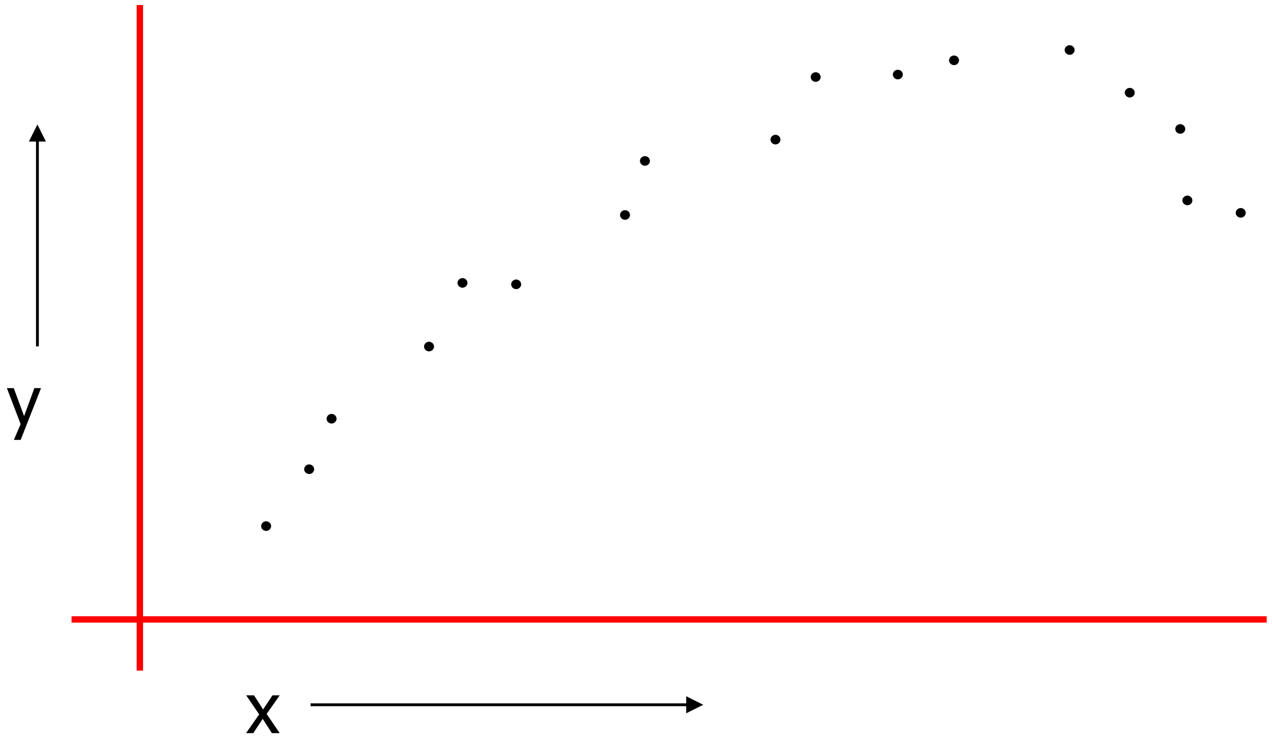
Test your understanding

Assuming regular regression trees, can you sketch a graph of the fitted function $y^{est}(x)$ over this diagram?



Test your understanding

Assuming linear regression trees, can you sketch a graph of the fitted function $y^{est}(x)$ over this diagram?



Copyright © 2001, 2003, Andrew W. Moore