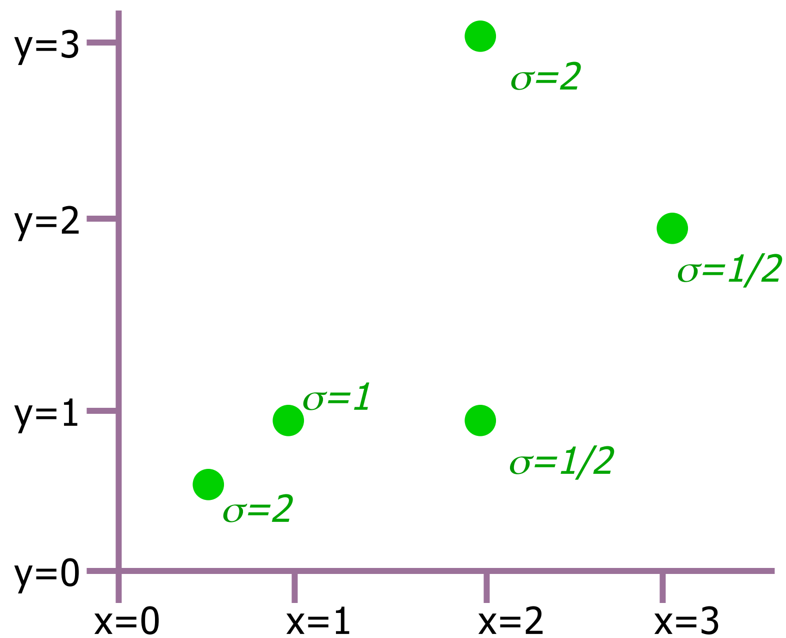
Heteroscedasticity...

Linear Regression with varying noise

Regression with varying noise

- Suppose you know the variance of the noise that was added to each datapoint.

x_i	y_i	σ_i^2
½	½	4
1	1	1
2	1	1/4
2	3	4
3	2	1/4



Assume $$y_i \sim N(wx_i, \sigma_i^2)$$

What's the MLE estimate of w?

MLE estimation with varying noise

$$\underset{w}{\operatorname{argmax}} \log p(y_1, y_2, \ldots, y_R \mid x_1, x_2, \ldots, x_R, \sigma_1^2, \sigma_2^2, \ldots, \sigma_R^2, w) =$$

$$\underset{w}{\operatorname{argmin}} \sum_{i=1}^{R} \frac{(y_i - wx_i)^2}{\sigma_i^2} =$$

Assuming independence among noise and then plugging in equation for Gaussian and simplifying.

$$\left(w \text{ such that } \sum_{i=1}^{R} \frac{x_i(y_i - wx_i)}{\sigma_i^2} = 0 \right) =$$

Setting dLL/dw equal to zero

$$\frac{\left(\sum_{i=1}^{R} \frac{x_i y_i}{\sigma_i^2} \right)}{\left(\sum_{i=1}^{R} \frac{x_i^2}{\sigma_i^2} \right)}$$

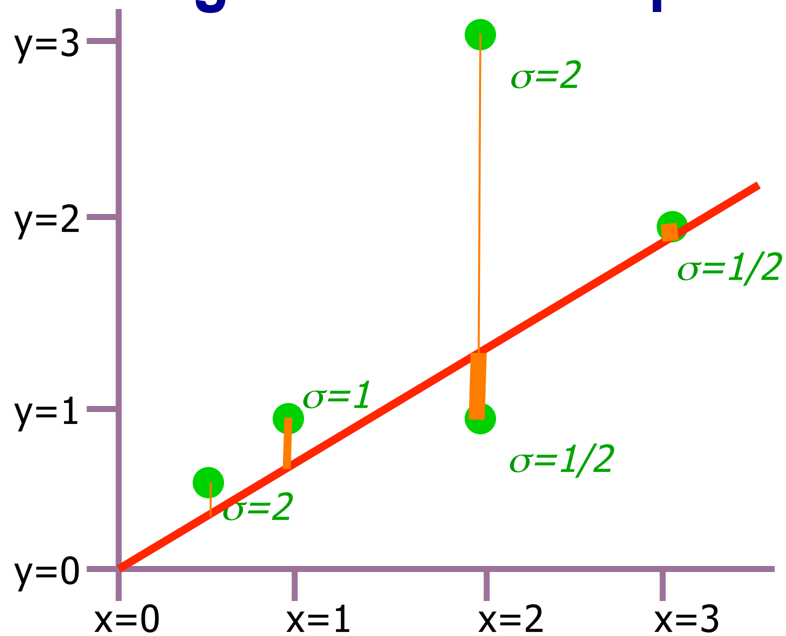
Trivial algebra

Copyright © 2001, 2003, Andrew W. Moore

This is Weighted Regression

- We are asking to minimize the weighted sum of squares

$$\underset{w}{\operatorname{argmin}} \sum_{i=1}^{R} \frac{(y_i - wx_i)^2}{\sigma_i^2}$$



where weight for i'th datapoint is $\frac{1}{\sigma_i^2}$

Copyright © 2001, 2003, Andrew W. Moore

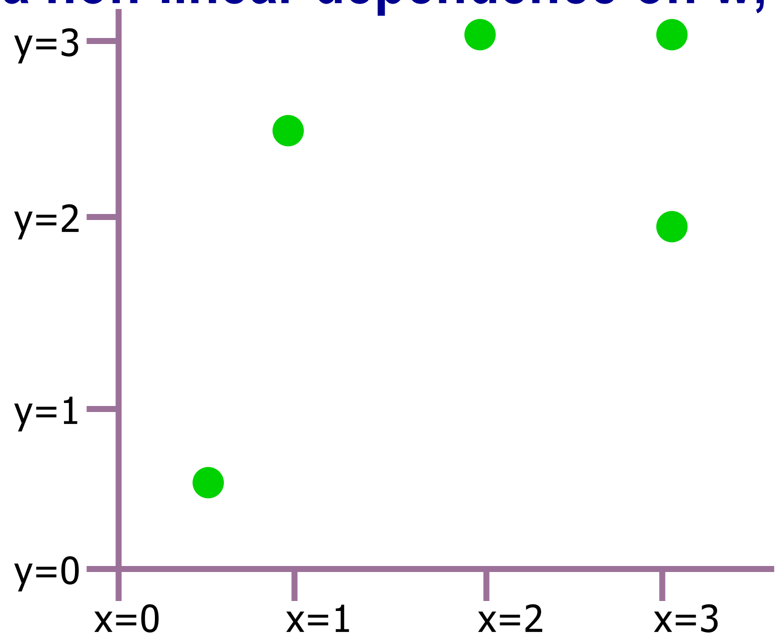
Non-linear Regression

Copyright © 2001, 2003, Andrew W. Moore

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?

Copyright © 2001, 2003, Andrew W. Moore

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

Copyright © 2001, 2003, Andrew W. Moore

Non-linear MLE estimation

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

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

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

Setting dLL/dw equal to zero

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

We' re down the algebraic toilet

So guess what we do?

Copyright © 2001, 2003, Andrew W. Moore

Non-linear MLE estimation

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

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

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

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

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?

Copyright © 2001, 2003, Andrew W. Moore

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
:	:	:

x=

3	2
1	1
:	:

y=

7
3
:

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

z=

1	3	2
1	1	1
:		:

y=

7
3
:

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

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

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

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

Copyright © 2001, 2003, Andrew W. Moore

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
:	:	:

x=

3	2
1	1
:	:

y=

7
3
:

$y_1=7$..

z=

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

y=

7
3
:

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

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

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

Copyright © 2001, 2003, Andrew W. Moore

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 $O(m^6)$

X_1	X_2
3	2
1	1
:	:

z=

1
1
:

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

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

Copyright © 2001, 2003, Andrew W. Moore

Q^{th}-degree polynomial Regression

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

x=

3	2
1	1
:	:

y=

7
3
:

z=

1	3	2	9	6	…
1	1	1	1	1	…
:					…

y=

7
3
:

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

m inputs, degree Q: how many terms?

= the number of unique terms of the form

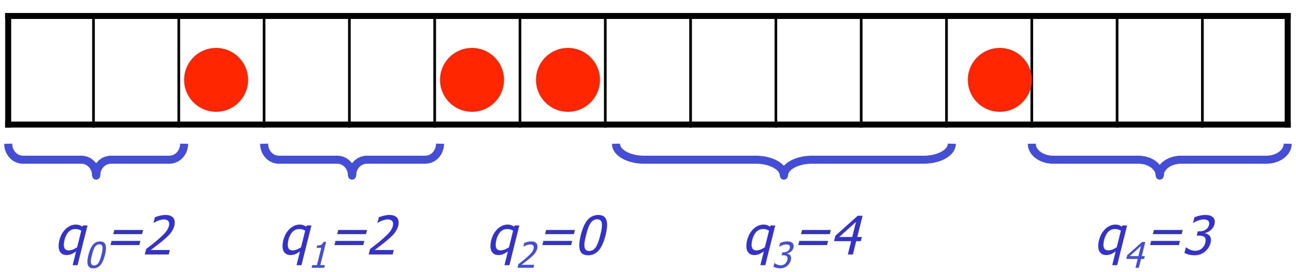
$$x_1^{q_1} x_2^{q_2} \ldots x_m^{q_m} \text{ where } \sum_{i=1}^{m} q_i \leq Q$$

= the number of unique terms of the form

$$1^{q_0} x_1^{q_1} x_2^{q_2} \ldots x_m^{q_m} \text{ where } \sum_{i=0}^{m} q_i = Q$$

= the number of lists of non-negative integers $[q_0, q_1, q_2, .. q_m]$ in which $\Sigma q_i = Q$

= the number of ways of placing Q red disks on a row of squares of length Q+m = (Q+m)-choose-Q



Q=11, m=4

Copyright © 2001, 2003, Andrew W. Moore

What we have seen

- **MLE with Gaussian noise is the same as minimizing the L_2 error**
  - Other noise models will give other loss functions
- **MLE with a Gaussian prior adds a penalty to the L_2 error, given Ridge regression**
  - Other priors will give different penalties
- **One can make nonlinear relations linear by transforming the features**
  - Polynomial regression
  - Radial Basis Functions (RBF) – will be covered later
  - Kernel regression (more on this later)

Copyright © 2001, 2003, Andrew W. Moore