Linear Regression

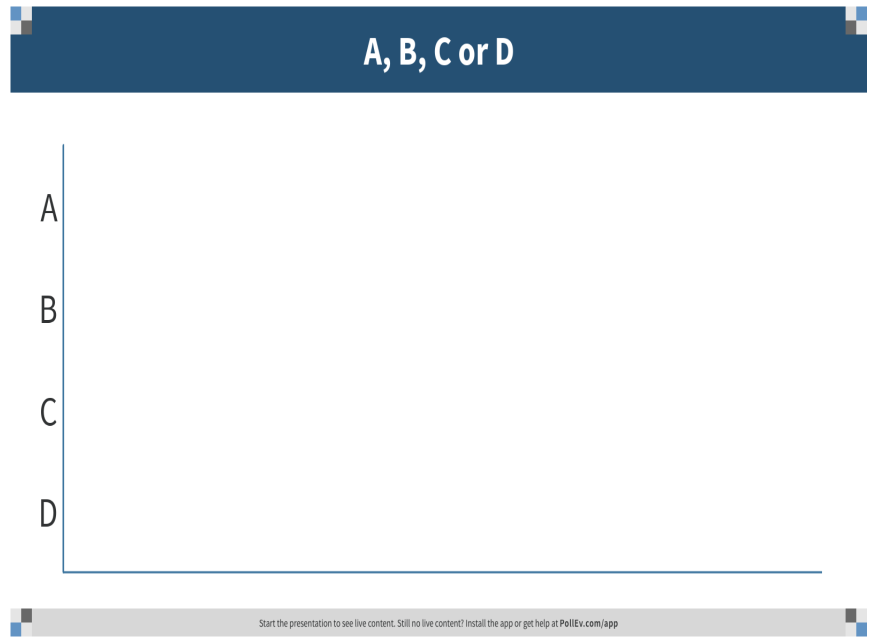
Lyle Ungar

Learning objectives

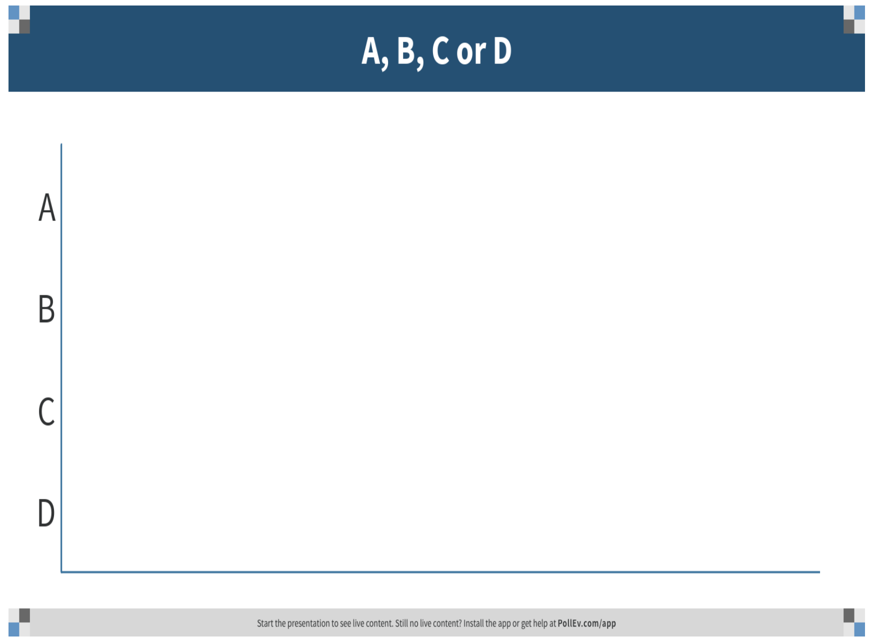
Be able to derive MLE & MAP regression and the associated loss functions

Recognize *scale invariance*

- **MLE estimates**
  - A) $\text{argmax}_\theta\ p(\theta|\mathbf{D})$
  - B) $\text{argmax}_\theta\ p(\mathbf{D}|\theta)$
  - C) $\text{argmax}_\theta\ p(\mathbf{D}|\theta)p(\theta)$
  - D) None of the above



- **MAP estimates**
  - A) $\arg\max_\theta \, p(\theta|\mathbf{D})$
  - B) $\arg\max_\theta \, p(\mathbf{D}|\theta)$
  - C) $\arg\max_\theta \, p(\mathbf{D}|\theta)p(\theta)$
  - D) None of the above



Consistent estimator

- A *consistent estimator* (or *asymptotically consistent estimator*) is an estimator — a rule for computing estimates of a parameter θ — having the property that as the number of data points used increases indefinitely, the resulting sequence of estimates converges in probability to the true parameter θ.

https://en.wikipedia.org/wiki/Consistent_estimator

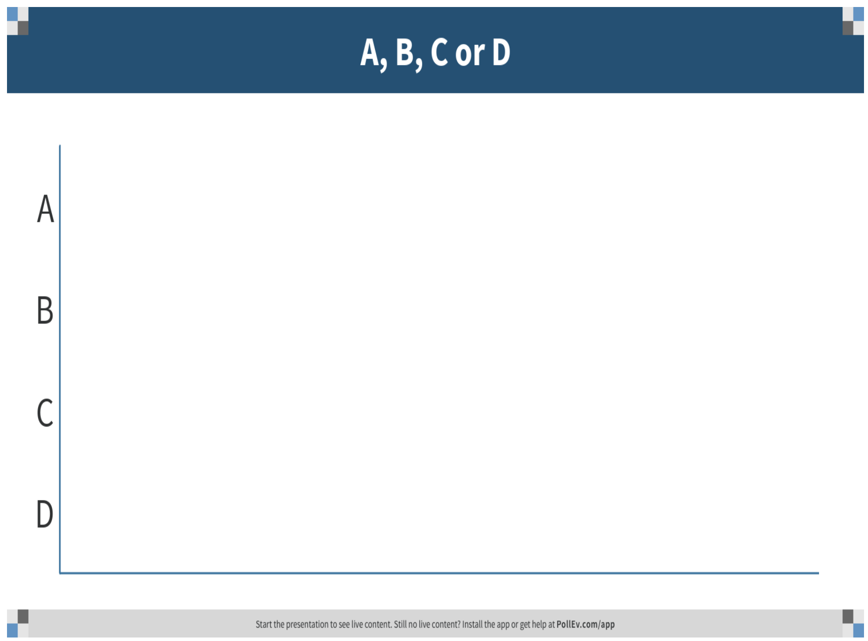
Which is consistent for our coin-flipping example?

A) MLE

B) MAP

C) Both

D) Neither



$P(D|\theta)$

$P(\theta|D) \sim P(D|\theta)P(\theta)$

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.

An introduction to regression

Mostly by Andrew W. Moore

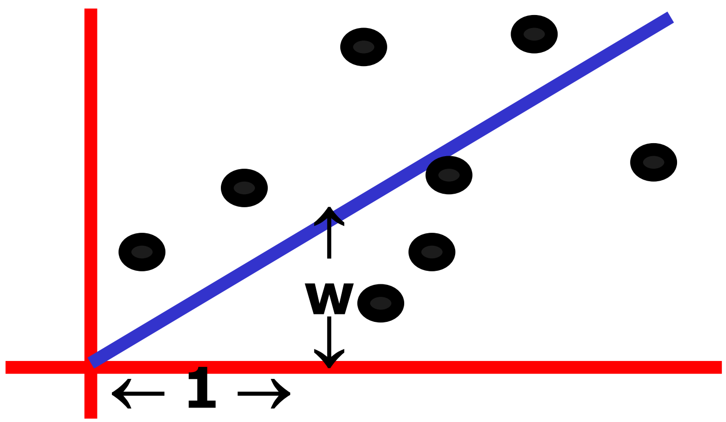
But with many modifications by Lyle Ungar

Two interpretations of regression

- **Linear regression**
  - $\hat{y} = \mathbf{w} \cdot \mathbf{x}$
- **Probabilistic/Bayesian (MLE and MAP)**
  - $y(\mathbf{x}) \sim N(\mathbf{w} \cdot \mathbf{x}, \sigma^2)$
  - **MLE:** $\mathrm{argmax}_{\mathbf{w}}\ p(\mathbf{D}|\mathbf{w})$ here: $\mathrm{argmax}_{\mathbf{w}}\ p(\mathbf{y}|\mathbf{w},\mathbf{X})$
  - **MAP:** $\mathrm{argmax}_{\mathbf{w}}\ p(\mathbf{D}|\mathbf{w})p(\mathbf{w})$
- **Error minimization**
  - $||\mathbf{y} - \mathbf{X}\mathbf{w}||_p^p + \lambda\ ||\mathbf{w}||_q^q$

Single-Parameter Linear Regression

Linear Regression



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 in x.

Simplest case: $\hat{y}(x) = wx$ for some unknown *w*.

Given the data, we can estimate *w*.

One parameter linear regression

Assume that the data is formed by

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

where…

- noise_i is independent $N(0, \sigma^2)$
- variance σ^2 is unknown

y(x) then **has a normal distribution with**

- mean wx
- variance σ^2

Bayesian Linear Regression

$p(y|w,x)$ is Normal(mean: wx, variance: σ^2)

$$y \sim N(wx, \sigma^2)$$

We have a data (x_1,y_1) (x_2,y_2) … (x_n,y_n)

We want to infer *w* from the data.

$$p(w|x_1, x_2, x_3,\ldots x_n, y_1, y_2\ldots y_n) = P(w|\mathbf{D})$$

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

Maximum likelihood estimation of *w*

MLE asks :

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

<=>

For what *w* is

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

<=>

For what *w* is $\prod_{i=1}^{n} p(y_i | w, x_i)$ 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

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

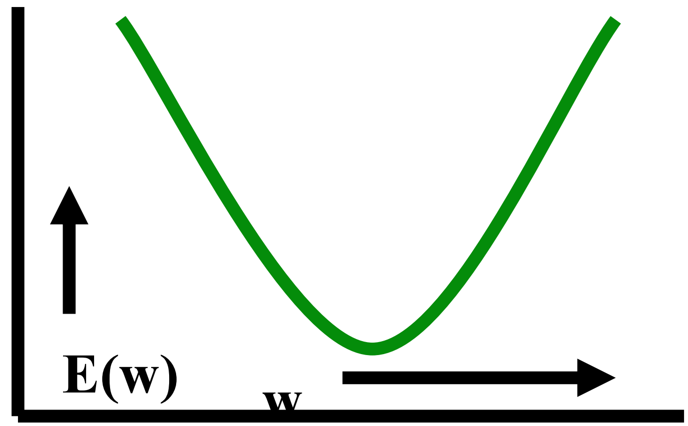
Result: MLE = L_2 error

- **MLE with Gaussian noise is the same as minimizing the L_2 error**

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

Linear Regression

The maximum likelihood *w* is the one that minimizes sum-of-squares of residuals



$$E = \sum_i (y_i - wx_i)^2$$

$$r_i = y_i - w\, x_i$$

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

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

Linear Regression

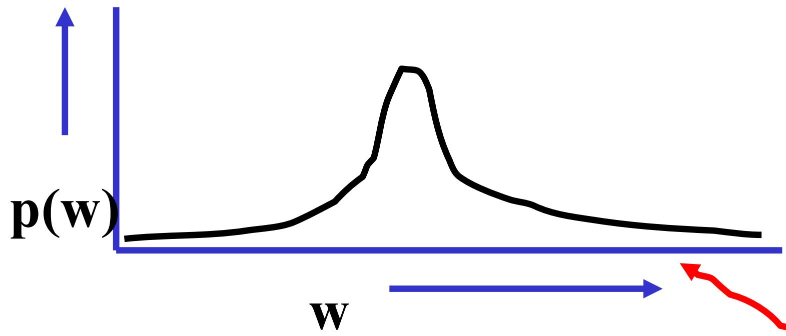
The sum of squares is minimized when

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

The maximum likelihood model is

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

We can use it for prediction



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

And predictions would have given a prob. distribution of expected output

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

But what about MAP?

- **MLE**

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

- **MAP**

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

But what about MAP?

- **MAP**

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

- **We assumed**
  - $y_i \sim N(w\, x_i, \sigma^2)$
- **Now add a prior assumption that**
  - $w \sim N(0, \gamma^2)$

For what w is

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

For what w is

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

For what w is

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

For what w is

$$\sum_{i=1}^{n} \left(y_i - wx_i\right)^2 + \left(\frac{\sigma w}{\gamma}\right)^2 \quad \text{minimized?}$$

Ridge Regression is MAP

- **MAP with a Gaussian prior on *w* is the same as minimizing the L_2 error plus an L_2 penalty on w**

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

- **This is called**
  - Ridge regression
  - Shrinkage
  - Tikhonov Regularization

Ridge Regression (MAP)

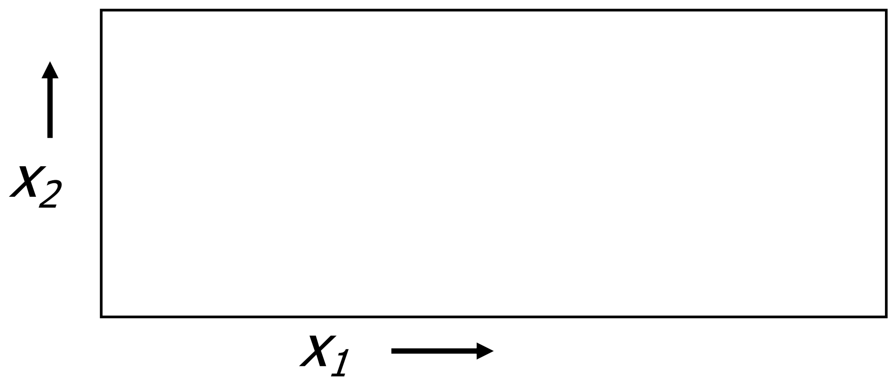
- **$w = x'y/(x'x + \lambda)$**

  $=(x'x + \lambda)^{-1} x'y$

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
$\vdots$	$\vdots$
$\mathbf{x}_n$	y_n

Multivariate Regression

Write matrix X and Y thus:

$$\mathbf{x} = \begin{bmatrix} \ldots\ldots\mathbf{x}_1\ldots\ldots \\ \ldots\ldots\mathbf{x}_2\ldots\ldots \\ \vdots \\ \ldots\ldots\mathbf{x}_n\ldots\ldots \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ & & \vdots & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

(n data points; Each input has p features)

The linear regression model assumes

$$\hat{y}(\boldsymbol{x}) = \boldsymbol{x} \cdot \boldsymbol{w} = w_1 x_1 + w_2 x_2 + \ldots. w_p x_p$$

Multivariate Regression

The maximum likelihood estimate (MLE) is

$$w = (X^T X)^{-1}(X^T y)$$

$X^T X$ is p x p

$X^T y$ is p x 1

Multivariate Regression

The MAP estimate is

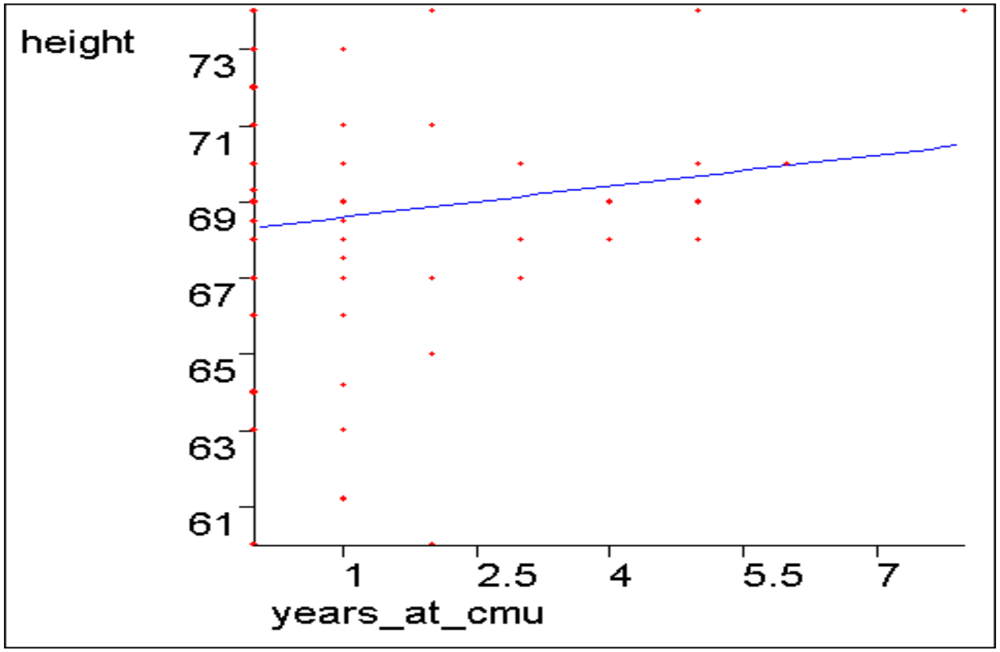
$$w = (X^T X + \lambda\, I)^{-1}(X^T y)$$

$X^T X$ is p x p

$X^T y$ is p x 1

What about a constant term?

Linear data usually 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: create a fake input "x_0" that is always 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

L_1 regression

OLS = L_2 regression minimizes

$p(y|w,x) \sim \exp(-||y-w^Tx||_2^2/2\sigma^2) \dashrightarrow \operatorname{argmin}_w ||\mathbf{y}-\mathbf{w}^T\mathbf{x}||_2^2$

L_1 regression:

$p(y|w,x) \sim \exp(-||y-w^Tx||_1/2\sigma^2) \dashrightarrow \operatorname{argmin}_w ||\mathbf{y}-\mathbf{w}^T\mathbf{x}||_1$

Scale Invariance

- **Rescaling *does not* affect decision trees or OLS**
  - They are scale invariant
- **Rescaling *does* affect Ridge regression**
  - Because it preferentially shrinks ‘large’ coefficients

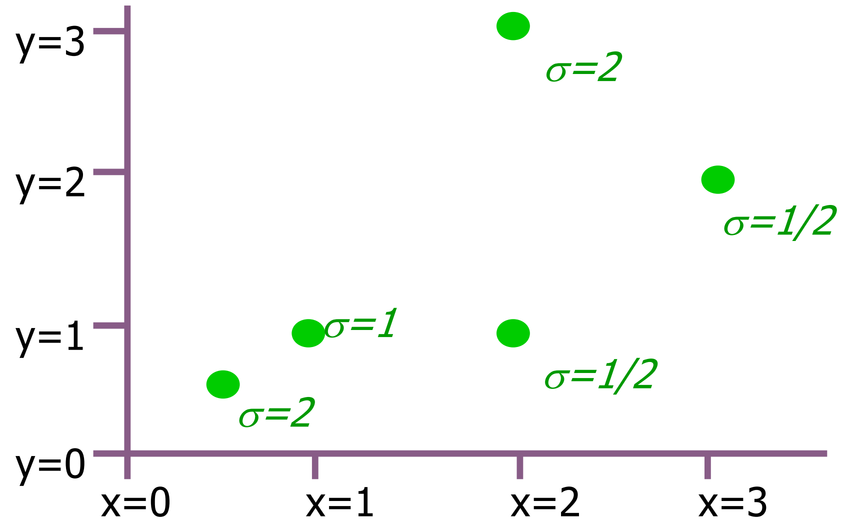
Heteroscedasticity...

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

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

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

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

Setting dLL/dw equal to zero

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

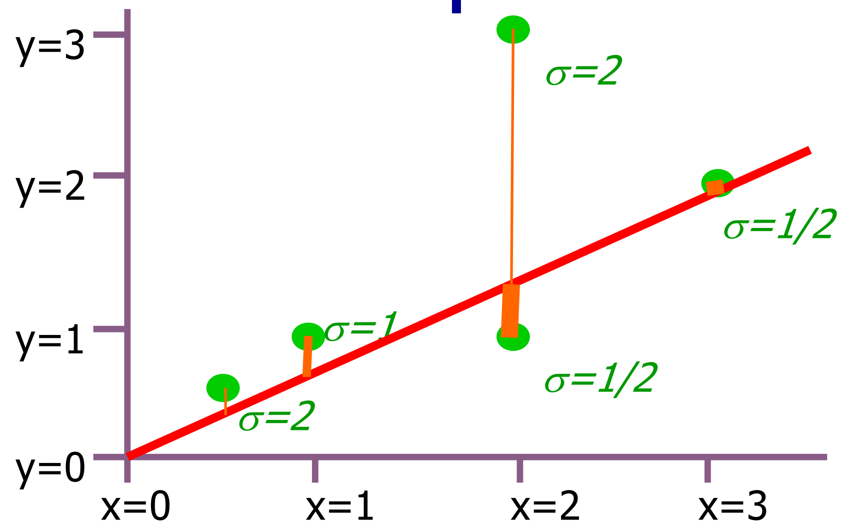
Trivial algebra

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

This is Weighted Regression

- We are minimizing the *weighted* sum of squares

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



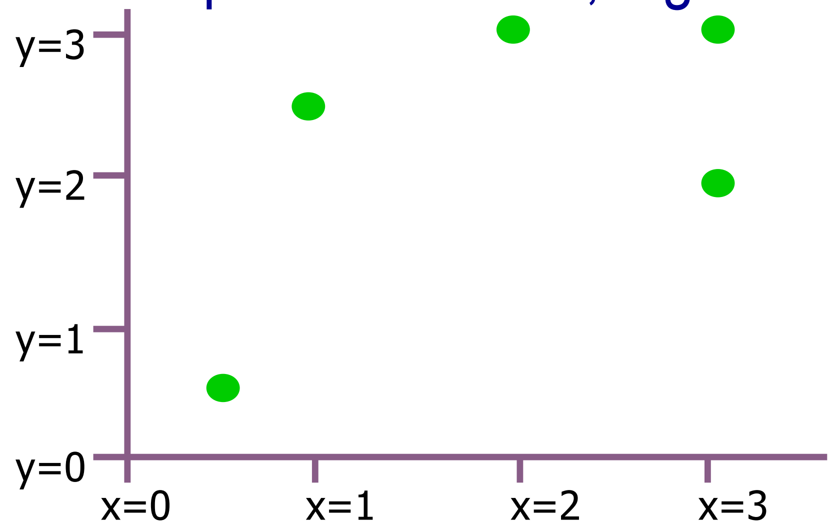
where the weight for i'th datapoint is $\frac{1}{\sigma_i^2}$

Nonlinear Regression

Nonlinear 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?

Nonlinear 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

Nonlinear MLE estimation

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

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

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?

Nonlinear MLE estimation

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

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

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

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

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

Also, special purpose statistical-optimization-specific tricks such as EM

We're down the algebraic toilet

So guess what we do?

What we have seen

- **MLE with Gaussian noise minimizes the L_2 error**
  - Other noise models will give other loss functions
- **MLE with a Gaussian prior gives Ridge regression**
  - Other priors will give different penalties
- **One can**
  - do nonlinear regression
  - make nonlinear relations linear by transforming the features