# An introduction to regression

**Mostly by Andrew W. Moore**

**But with modifications by Lyle Ungar**



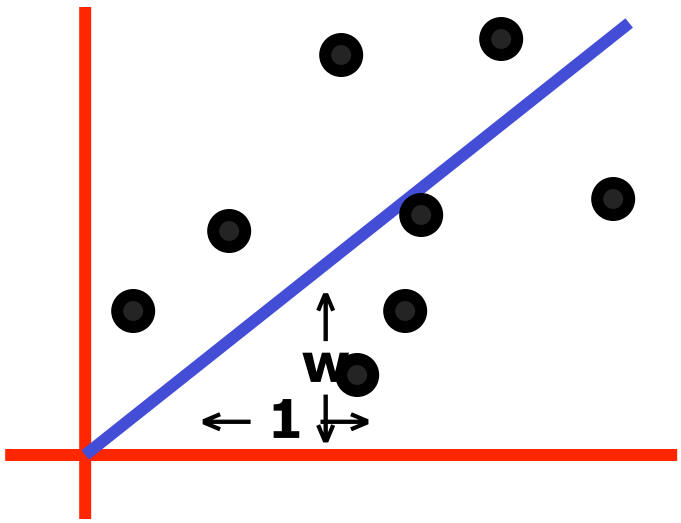

Copyright © Andrew W. Moore

# Two interpretations of regression

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

# 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 $\sigma^2$

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

- mean $wx$
- variance $\sigma^2$

# Bayesian Linear Regression

$$p(y|w,x) = \text{Normal (mean } wx\text{, var } \sigma^2)$$

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

We have a set of data $(x_1,y_1)\ (x_2,y_2)\ \ldots\ (x_n,y_n)$

We want to infer $w$ from the data.

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

- 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*

**MLE asks the :**

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

**<=>**

**For what *w* is**

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

**<=>**

**For what *w* is**

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

For what *w* is

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

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

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

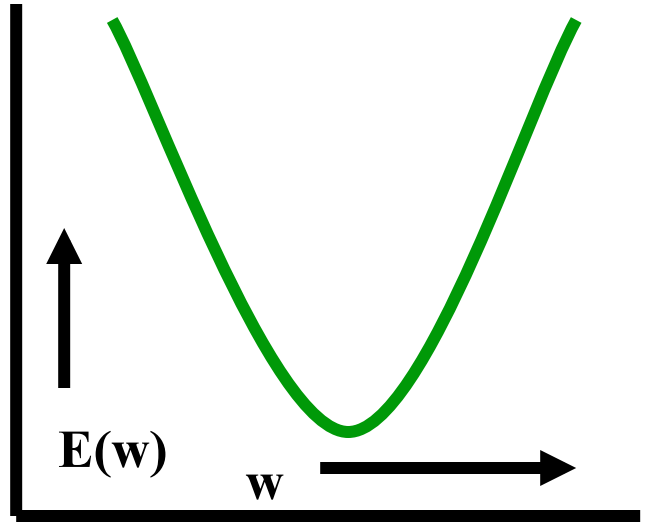
# First result

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

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

# Linear Regression

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




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

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

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

$$\mathrm{Out}(x) = wx$$

We can use it for prediction

# Linear Regression

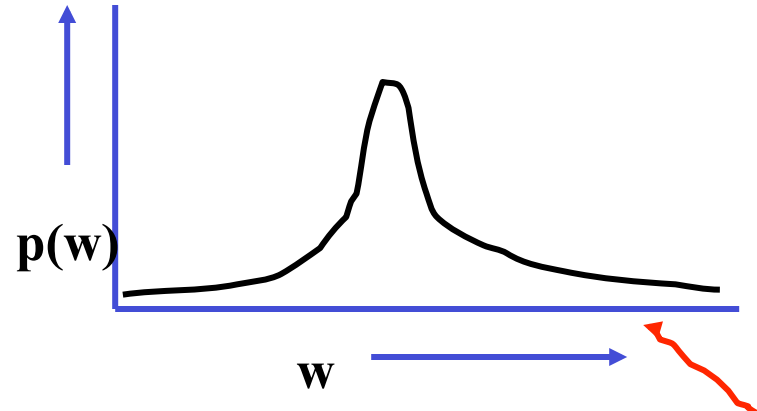
**Easy to show the sum of squares is minimized when**

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

The maximum likelihood model is

$$\mathrm{Out}(x) = wx$$

We can use it for prediction




**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.

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

For what *w* is

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

For what *w* is

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

For what *w* is

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

For what *w* is

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

# Second result

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

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

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