



Linear Regression & Least Squares

Maximum Likelihood & LS

Underconstrained LS

Not only lines!

Overfitting



Maximum Likelihood and LS

- Given a particular set of regression parameters w , what is the probability for the data set $\{x(t), y(t)\}$ to occur?
- The probability of the data given the parameters is the *likelihood* of the parameters given the data.
- Under certain assumptions, least squares finds the parameters corresponding to maximum likelihood!



Maximum Likelihood and LS

- Assumptions:

- Each observed value $y(t)$ has a measurement error $\varepsilon(t)$

$$y(t) = Xw + \varepsilon(t)$$

- The errors $\varepsilon(t)$ are independent random variables, normally distributed with mean of 0, and all have the same standard deviation σ .

$$P \propto \prod_{t=1}^m \left\{ \exp \left[-\frac{1}{2} \left(\frac{y(t) - w^T x(t)}{\sigma} \right)^2 \right] \Delta y \right\}$$



Maximum Likelihood and LS

- Maximizing the probability P:

$$P \propto \prod_{t=1}^m \left\{ \exp \left[-\frac{1}{2} \left(\frac{y(t) - w^T x(t)}{\sigma} \right)^2 \right] \Delta y \right\}$$

- is equivalent to minimizing $-\log(P)$:

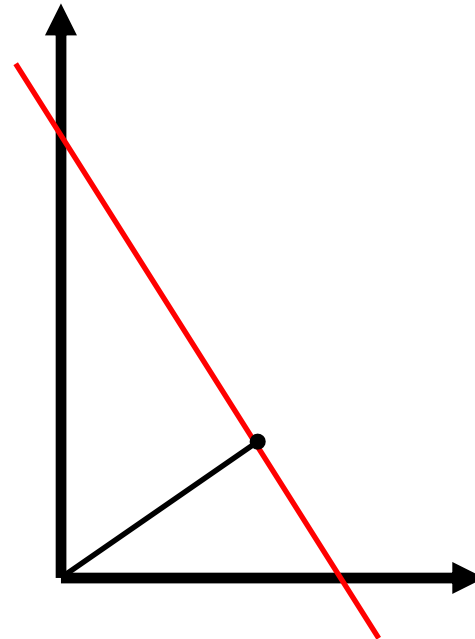
$$\left[\sum_{t=1}^m \frac{(y(t) - w^T x(t))^2}{2\sigma^2} \right] - m \log \Delta y$$

Underconstrained LS

- What if we have more unknowns than equations ($m \ll n$)?

- Example:
$$30x_1 + 10x_2 = 400$$

- Multiple solutions:
 $(0, 40), (40/3, 0)$

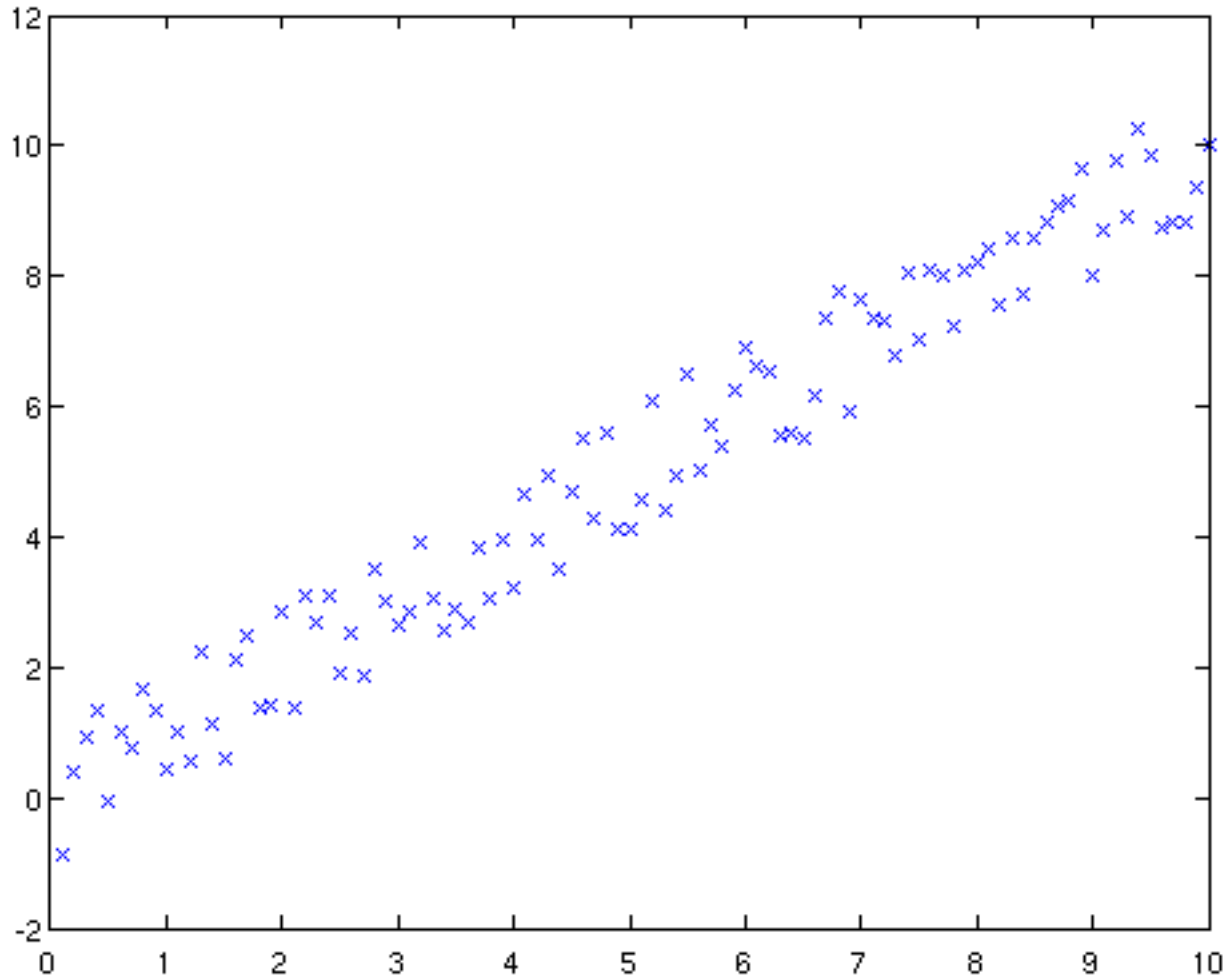




Underconstrained LS

- In the general (underconstrained) case, let $\text{Null}(A) = \{v \mid Av = 0\}$.
- If y is a solution to $Ax = b$, then so is any vector $(y+z)$, where $z \in \text{Null}(A)$.
- A LS solution perpendicular to $\text{Null}(A)$, must have the form: $A^T u$
- Substitute and solve $AA^T u = b$
- So the LS solution is $x = A^T (AA^T)^{-1} b$

Fitting in 1D: some noisy data





Fitting in 1D

line through origin:

$$\begin{bmatrix} x(1) \\ x(2) \\ \vdots \\ x(m) \end{bmatrix} [w] = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(m) \end{bmatrix}$$

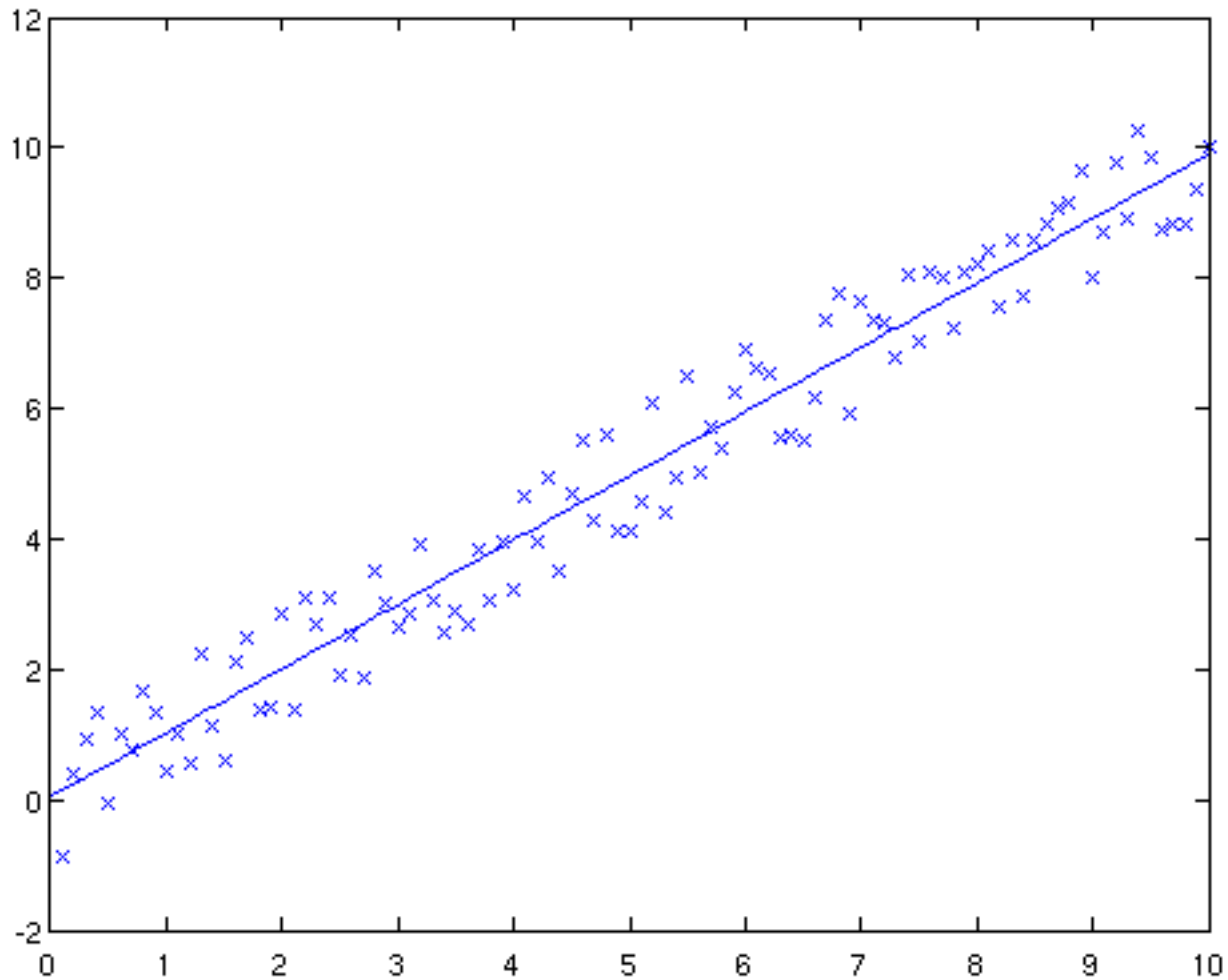
general line:

$$\begin{bmatrix} 1 & x(1) \\ 1 & x(2) \\ \vdots & \vdots \\ 1 & x(m) \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(m) \end{bmatrix}$$

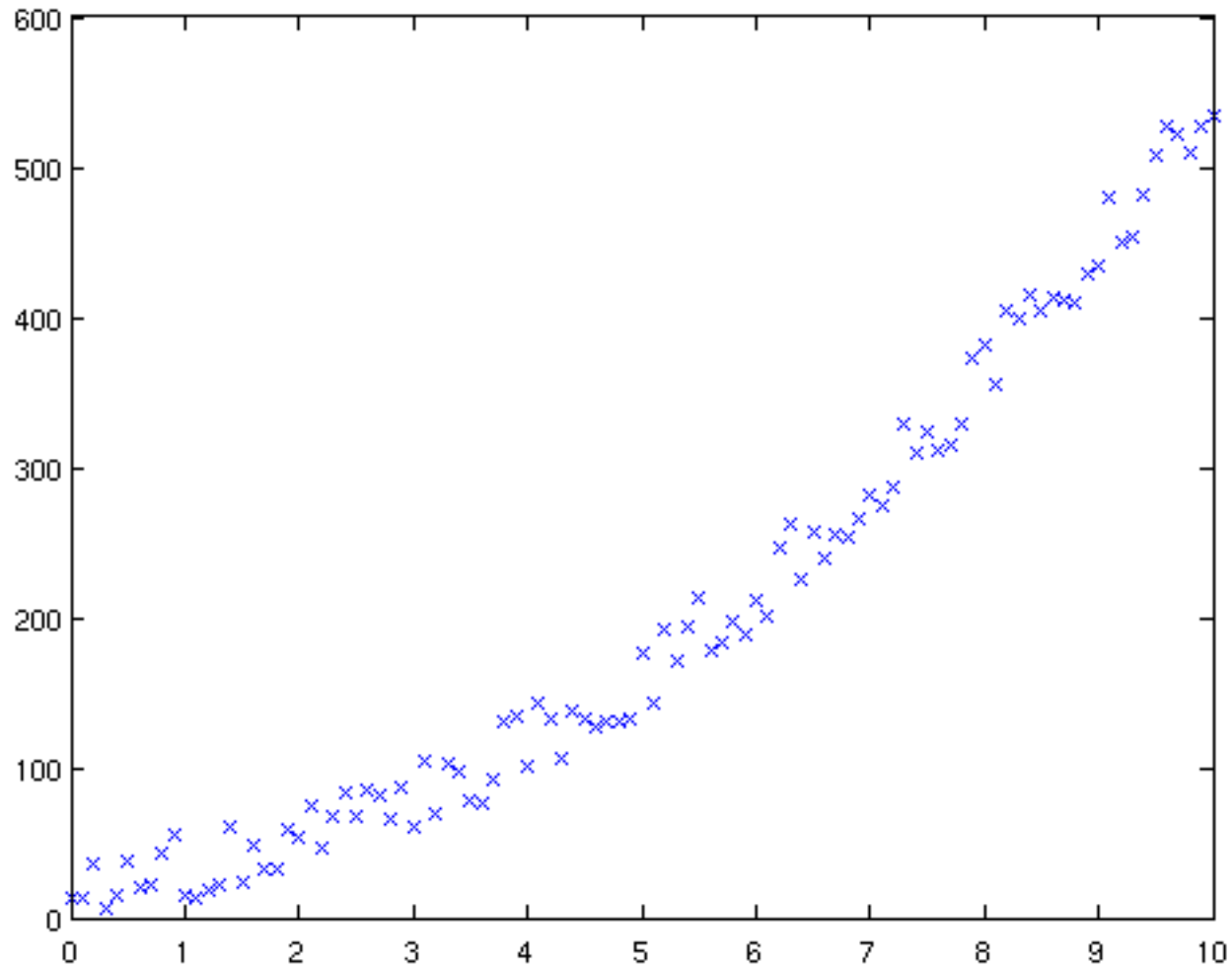
fitting a polynomial of degree n:

$$\begin{bmatrix} 1 & x(1) & \cdots & x(1)^n \\ 1 & x(2) & \cdots & x(2)^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x(m) & \cdots & x(m)^n \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(m) \end{bmatrix}$$

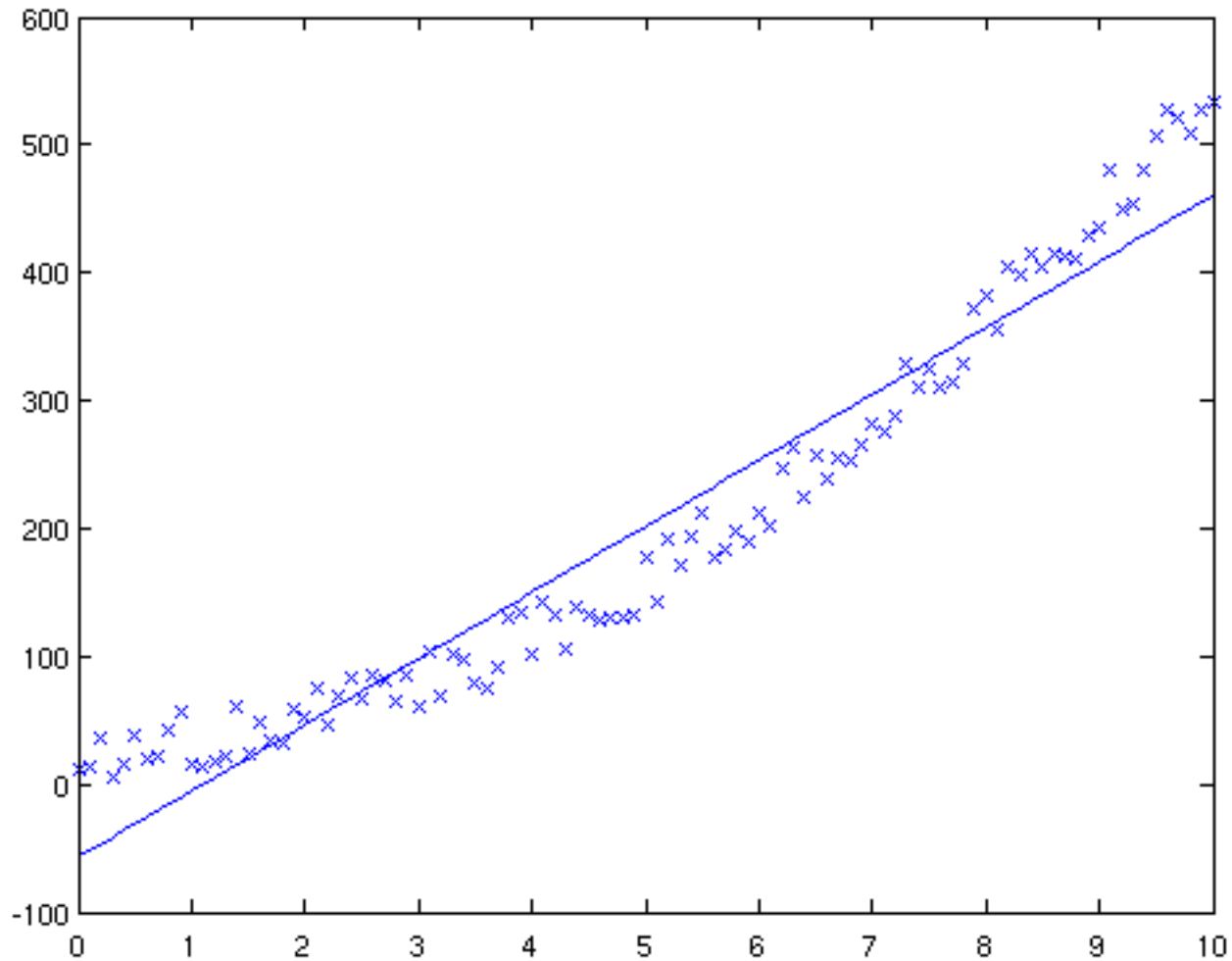
Least squares line fit



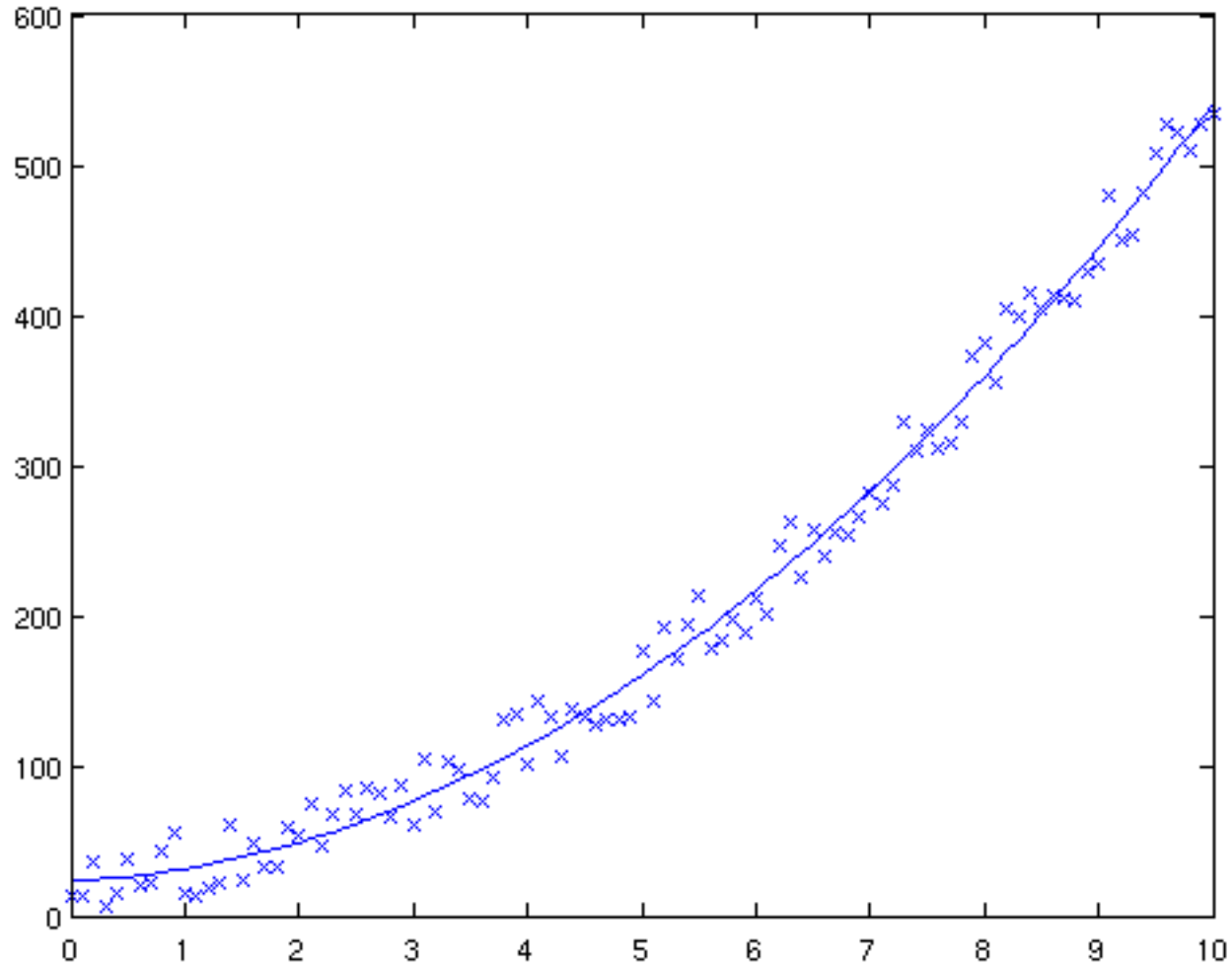
More noisy data



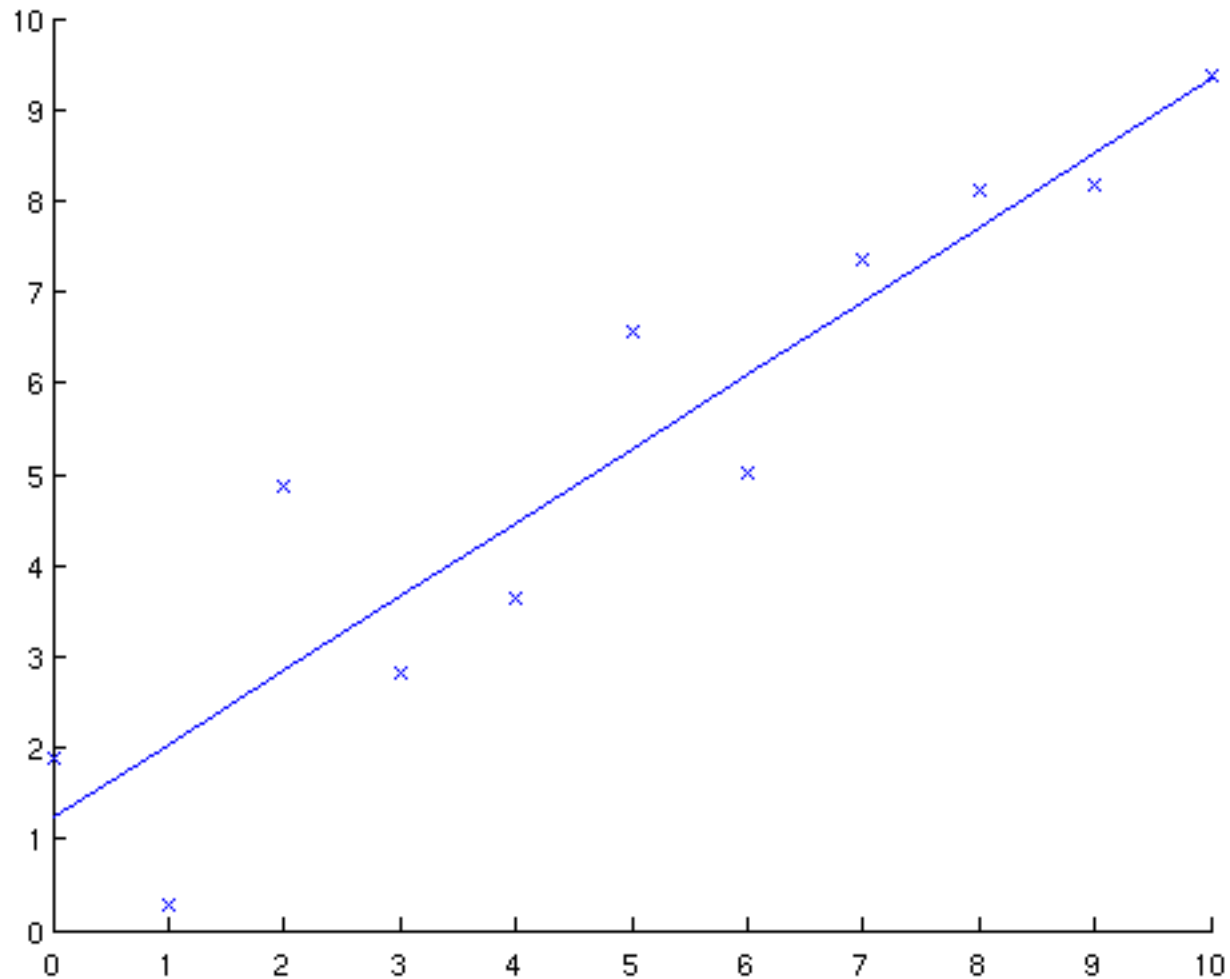
Least squares line fit



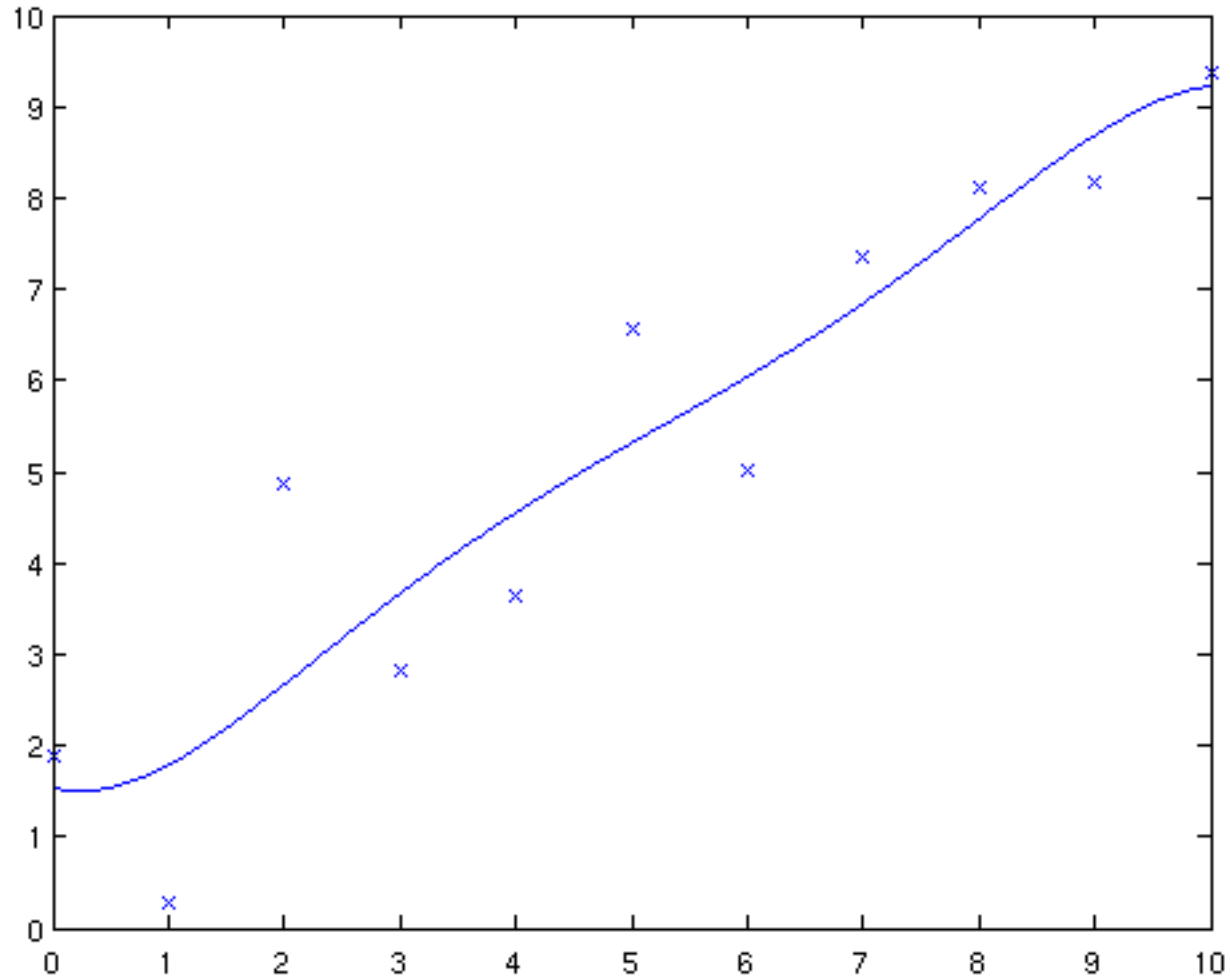
Least squares polynomial fit



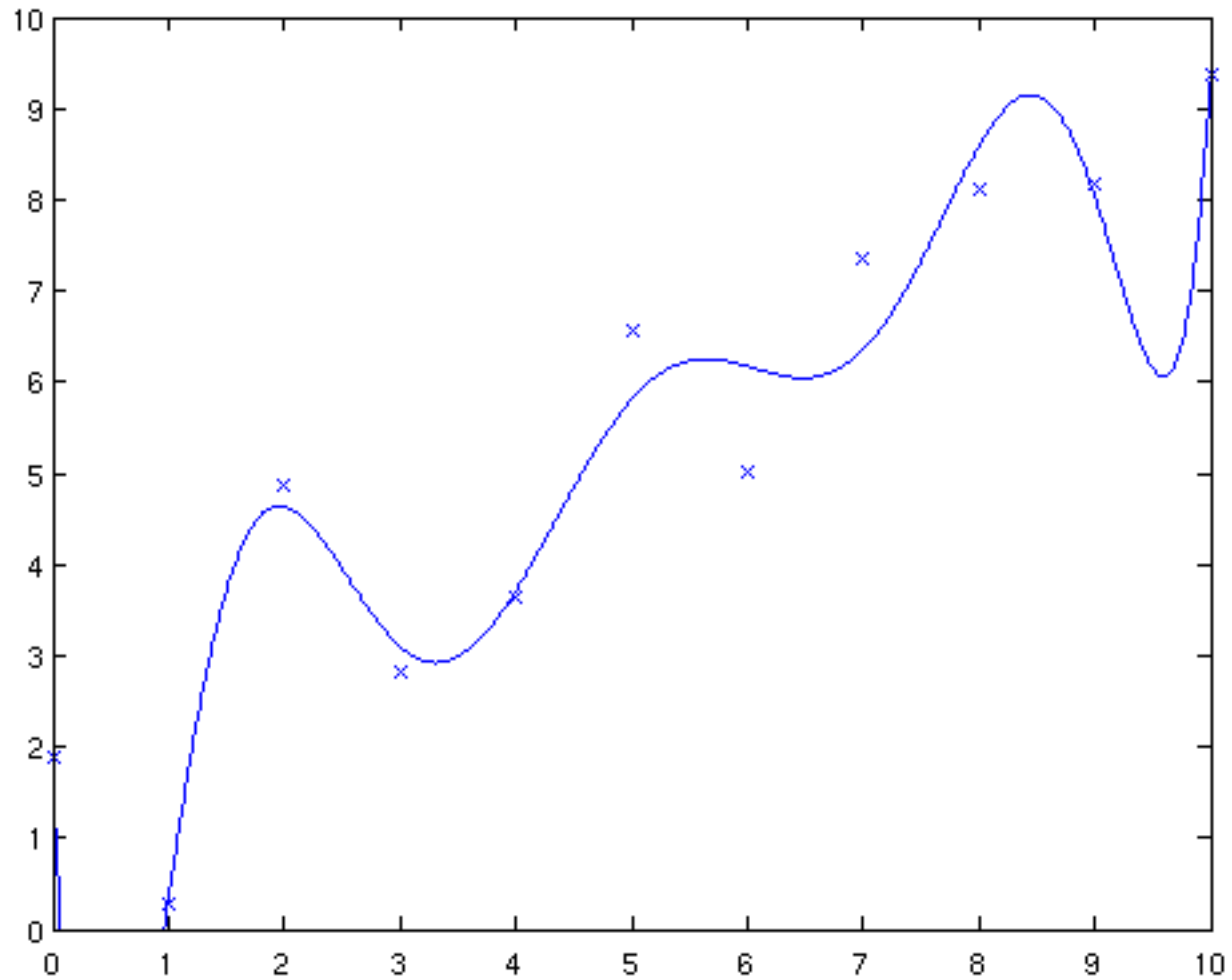
1st degree polynomial



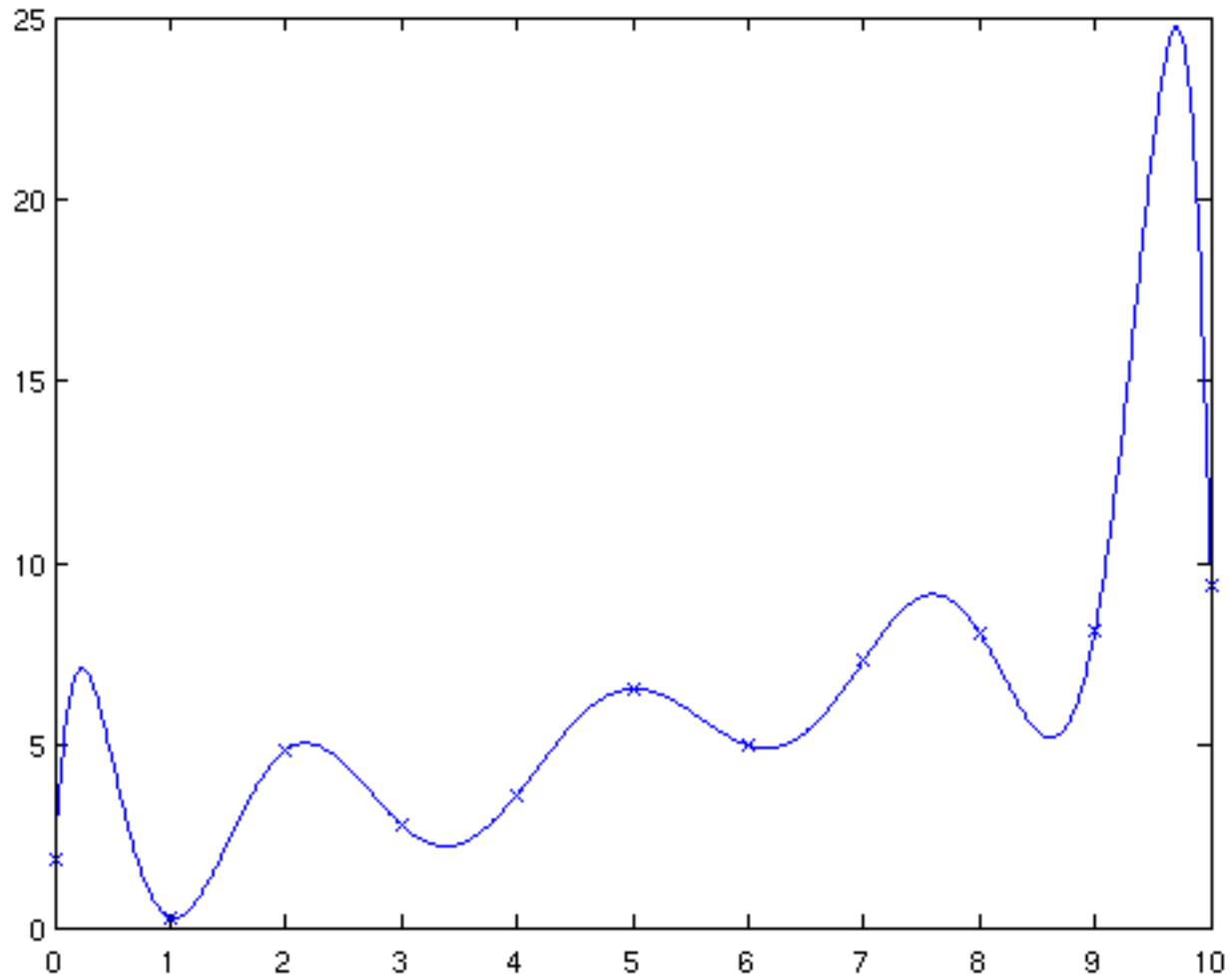
5th degree polynomial



8th degree polynomial



10th degree polynomial





A slight generalization

- The polynomial fit in 1D is an example of a generalized version of linear regression, where instead of using the input variable directly we use a vector valued function of the input variables:

$$\begin{bmatrix} f_1(x(1)) & f_2(x(1)) & \cdots & f_n(x(1)) \\ f_1(x(2)) & f_2(x(2)) & \cdots & f_n(x(2)) \\ \vdots & \vdots & \ddots & \vdots \\ f_1(x(m)) & f_2(x(m)) & \cdots & f_n(x(m)) \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(m) \end{bmatrix}$$



A Little Bit of History

- Sir Francis Galton F.R.S. (1822 - 1911)
 - Half cousin of Charles Darwin
 - Invented the use of the regression line
 - First to observe “regression towards the mean”
- Regression towards the mean:
 - the height of a child tends to be closer to the mean than the average of the heights of his/her parents
 - While it’s tempting to believe a scientific discovery has been made, simple statistics is at play here.



Non-linear regression

- Infer the parameters θ of a non-linear function of the inputs (independent variables) $\{x(t)\}$ that best predict the outputs (dependent variables) $\{y(t)\}$:

$$y = f(x, \theta) + \varepsilon$$

- In general, finding the parameters θ involves solving a non-linear optimization problem.



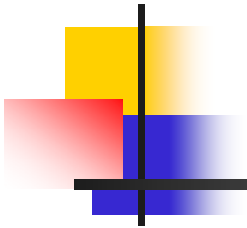
Linearization

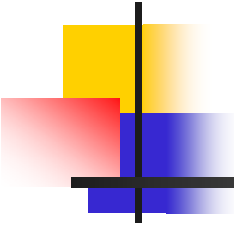
- Sometimes a non-linear problem may be linearized by a suitable transformation.
- Example:

$$y = ae^{bx}$$

$$\ln(y) = \ln(a) + bx$$

Conditioning of Linear Systems





Conditioning – An Example

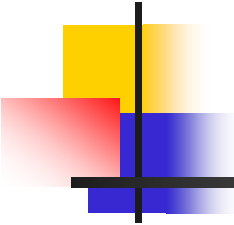
$$\begin{bmatrix} 600 & 800 \\ 30,001 & 40,002 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 200 \\ 10,001 \end{bmatrix}$$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

exact

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1.013887 \\ 1.010416 \end{bmatrix}$$

computed



Conditioning – An Example

$$\begin{bmatrix} 600 & 800 \\ 30,001 & 40,002 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 200 \\ 9,999 \end{bmatrix}$$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$$

exact



Conditioning – An Example

$$\begin{bmatrix} .3 & .4 \\ .30001 & .40002 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} .1 \\ .10001 \end{bmatrix}$$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \quad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -0.987987 \\ 0.990990 \end{bmatrix}$$

exact

computed



Condition Number

- The conditioning of a linear system $Ax = b$ may be quantitatively measured by the condition number of the matrix A :

$$\kappa(A) = \|A\| \|A^{-1}\|$$

$$1 = \|I\| = \|AA^{-1}\| \leq \|A\| \|A^{-1}\|$$



Condition Number

$$A = \begin{bmatrix} 600 & 800 \\ 30,001 & 40,002 \end{bmatrix}$$

$$A^{-1} \approx \begin{bmatrix} 100 & -2.0 \\ -75 & 1.5 \end{bmatrix}$$

$$\kappa_1(A) = \|A\|_1 \|A^{-1}\|_1 \approx 7 \times 10^6$$



Condition Number

- Let $A(x + \delta x) = b + \delta b$ where δb is the error in b , and δx is the resulting error in the solution x . Then:

$$\frac{\|\delta x\|}{\|x\|} \leq \kappa(A) \frac{\|\delta b\|}{\|b\|}$$

- Proof: next slide...



Condition Number

$$A(x + \delta x) = b + \delta b$$

$$Ax + A\delta x = b + \delta b$$

$$A\delta x = \delta b$$

$$\delta x = A^{-1}\delta b$$

$$\|\delta x\| \leq \|A^{-1}\| \|\delta b\|$$



Condition Number

$$\frac{\|\delta x\| / \|x\|}{\|\delta b\| / \|b\|} = \frac{\|A^{-1} \delta b\| / \|x\|}{\|\delta b\| / \|Ax\|} = \frac{\|Ax\|}{\|x\|} \frac{\|A^{-1} \delta b\|}{\|\delta b\|}$$

$$\frac{\|\delta x\| / \|x\|}{\|\delta b\| / \|b\|} \leq \frac{\|A\| \|x\| \|A^{-1}\| \|\delta b\|}{\|x\| \|\delta b\|} = \|A\| \|A^{-1}\|$$

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\| \|A^{-1}\| \frac{\|\delta b\|}{\|b\|}$$



Condition Number

- Let \tilde{x} denote a solution to the linear system $Ax = b$, and let $r = b - A\tilde{x}$ denote the residual. Then:

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leq \kappa(A) \frac{\|r\|}{\|b\|}$$

- Proof: follows from previous result.



Condition Number

- Let $(A + \delta A)(x + \delta x) = b$ where δA is the error in A , and δx is the resulting error in the solution x . Then:

$$\frac{\|\delta x\|}{\|x\|} \leq \kappa(A) \frac{\|\delta A\| / \|A\|}{1 - \|\delta A\| \|A^{-1}\|}$$

- Proof: next slide...



Condition Number

$$(A + \delta A)(x + \delta x) = b$$

$$Ax + A\delta x + \delta A(x + \delta x) = b$$

$$A\delta x = -\delta A(x + \delta x)$$

$$\delta x = -A^{-1}\delta A(x + \delta x)$$

$$\|\delta x\| \leq \|A^{-1}\| \|\delta A\| (\|x\| + \|\delta x\|)$$

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\|A^{-1}\| \|\delta A\|}{1 - \|A^{-1}\| \|\delta A\|} = \kappa(A) \frac{\|\delta A\| / \|A\|}{1 - \|A^{-1}\| \|\delta A\|}$$