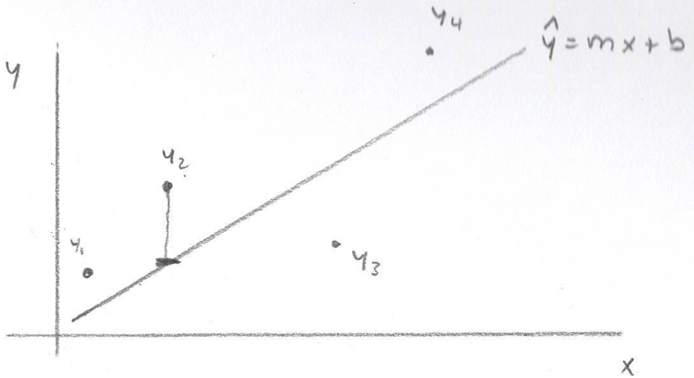
Linear Regression



n observations of y

y ' income

x : years of college
(explanatory variable)

Q: (1) degree of correlation ?

(2) best (linear) fit ?

SSE: Sum squared error

$$\sum(\hat{y}_i - y_i)^2 = \|\vec{\hat{y}} - \vec{y}\|^2$$ ← norm

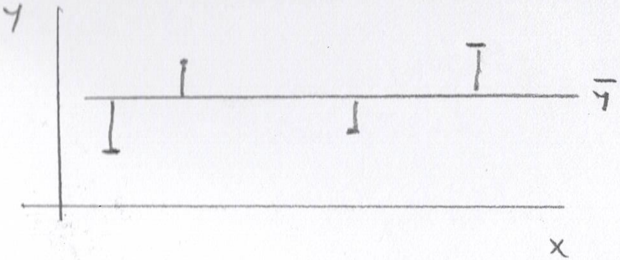
$$\|\vec{x}\| = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}$$

SSE = 0 ⇒ $\hat{y}_i = y_i$ for all i ⇒ perfect fit

SST: sum squared total "error if using mean as predictor"

$$\sum(y_i - \bar{y})^2$$

↖ mean of all observations



SST = 0 => all values equal mean

$$R^2 = 1 - \frac{SSE}{SST}$$ correlation coefficient

SSE = 0 => perfect fit $R^2 = 1$

SSE = SST => no correlation $R^2 = 0$

"no linear model is better than mean"

What is the best linear model?

$\hat{y} = mx + b$ predictor model

$$\left.\begin{aligned} \hat{y}_1 &= m x_1 + b \\ \hat{y}_2 &= m x_2 + b \\ &\vdots \\ \hat{y}_n &= m x_n + b \end{aligned}\right\} \quad \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} b \\ m \end{bmatrix}$$

$$\hat{Y} = A\beta$$

"best fit" $\Rightarrow$ minimize SSE

minimize $\|\hat{y} - y\|^2$

minimize $\|A\beta - y\|^2$ $\Rightarrow$ find β that minimizes SSE

steps $\dfrac{\partial SSE}{\partial \beta} = 0$ $\Rightarrow$ solve for β

$$SSE = \| A\beta - y \|^2$$

$$= (A\beta - y)^T (A\beta - y)$$

$$= (\beta^T A^T - y^T)(A\beta - y)$$

$$= \beta^T A^T A\beta - y^T A\beta - \underbrace{\beta^T A^T y}_{(A\beta)^T y} + y^T y$$

$(A\beta)^T y$

$y^T A\beta$) transpose

$$= \beta^T A^T A\beta - 2y^T A\beta + y^T y$$

$$\frac{\partial SSE}{\partial \beta} = ?$$

---

derivative rules

$$\frac{\partial \vec{y}}{\partial \vec{x}} \overset{def}{=} \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_n} \\ \vdots & & \\ \frac{\partial y_1}{\partial x_m} & \cdots & \frac{\partial y_n}{\partial x_m} \end{bmatrix}$$

useful identities

$$\|x^2\| = \Sigma x_i^2 = x^T x$$

$$\|x\| = \sqrt{\Sigma x_i}$$

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} [x_1 \ldots x_n] = x_1^2 + \ldots x_n^2$$

$$(A^T)^T = A$$

$$(AB)^T = B^T A$$

$$(A+B)^T = A^T + B^T$$

$$SSE = \beta^T A^T A \beta - 2 y^T A \beta + y^T y$$

$$\frac{\partial SSE}{\partial \beta} = 2 A^T A \beta - 2 A^T y + 0 = 0$$

$$A^T A \beta = A^T y$$

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

$$\boxed{\beta = (A^T A)^{-1} A^T y}$$

note: $A^T A$ is always invertible if column vectors are linearly independent

Common matrix derivatives

$\vec{y}$	$\frac{\partial \vec{y}}{\partial \vec{x}}$
$A\vec{x}$	A
$x^T A$	A^T
$x^T x$	$2x$
$x^T A^T A x$	$2 A^T A x$
$y^T A x$	$A^T y$

Linear regression in practice

x	y
1	-0.2
2	2.2
3	3.5
⋮	⋮
8	12

n explanatory variable — regressed variable

column of 1 — column of x_i — vector of y

$$\underset{n\times 2}{A} = \begin{bmatrix} 1 & -0.2 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & 8 \end{bmatrix} \qquad y = \begin{bmatrix} -0.2 \\ 2.2 \\ \vdots \\ 12 \end{bmatrix}$$

$$\beta = inv(A^T A)\, A\, y = \begin{bmatrix} m \\ b \end{bmatrix}$$

defined in any language

may need some Linear Algebra package

$$SSE = norm(Ax - y)^2$$

$$SST = norm(y - mean(y))^2$$

$$R^2 = 1 - \frac{SSE}{SST}$$

may need Statistics package

Multilinear Regression

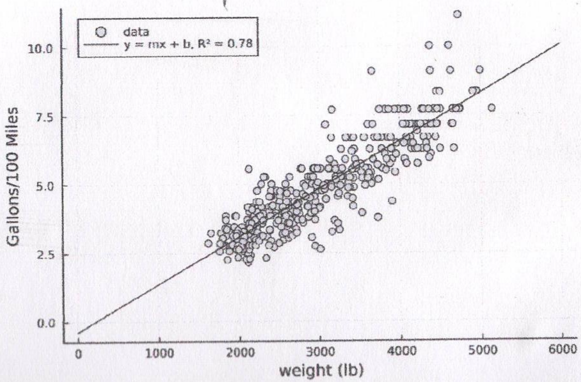
base case $y = \frac{1}{mpg}$ (Gallons per 100 miles)

$x = \text{weight}$

$y = mx + b$

$R^2 = 0.79$

	mpg	cylinders	displacement	hp	weight	acceleration	year	mo
1	18.0	8	307.0	130.0	3504.0	12.0	70	"chevrolet ch
2	15.0	8	350.0	165.0	3693.0	11.5	70	"buick skylarl
3	18.0	8	318.0	150.0	3436.0	11.0	70	"plymouth sat
4	16.0	8	304.0	150.0	3433.0	12.0	70	"amc rebel ss
5	17.0	8	302.0	140.0	3449.0	10.5	70	"ford torino"
6	15.0	8	429.0	198.0	4341.0	10.0	70	"ford galaxie
7	14.0	8	454.0	220.0	4354.0	9.0	70	"chevrolet im
8	14.0	8	440.0	215.0	4312.0	8.5	70	"plymouth fur
9	14.0	8	455.0	225.0	4425.0	10.0	70	"pontiac cata
10	15.0	8	390.0	190.0	3850.0	8.5	70	"amc ambassad
⋮ more								
392	31.0	4	119.0	82.0	2720.0	19.4	82	"chevy s-10"



"Cars" Dataset

Multivariable model

$$y = c_1 x_1 + c_2 x_2 + c_3 x_3 + c_4 \leftarrow \text{"intercept"}$$

x_1: weight

x_2: horsepower

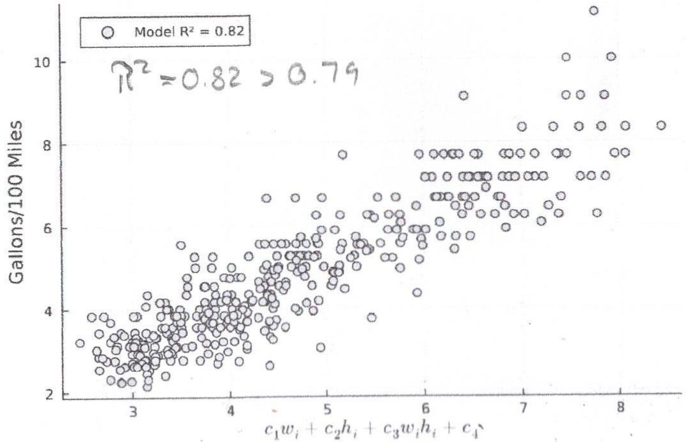
x_3: weight × horsepower

$$\hat{y} = A \beta$$

$$\begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{21} & x_{31} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{1n} & x_{2n} & x_{3n} & 1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix}$$

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

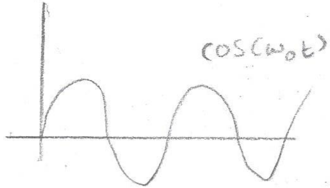
adding variables increases explanatory power

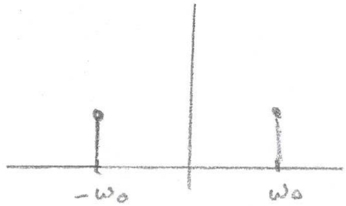


$R^2 = 0.82 > 0.79$

Integral Equations (Domain Transformations)

Fourier Transform $\quad F(\omega) = \int_{-\infty}^{\infty} \exp(-i\omega t)\, f(t)\, dt$



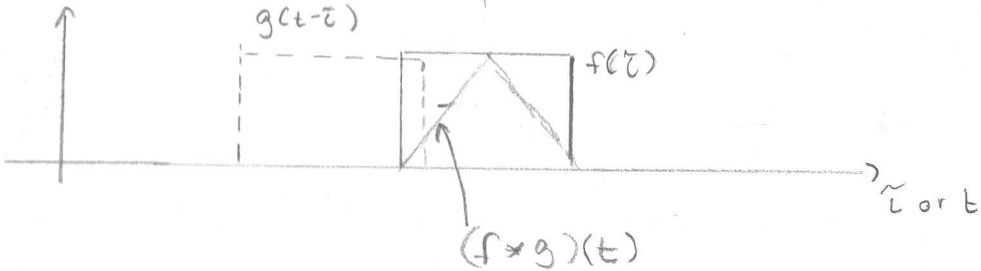


t-domain $\longleftrightarrow$ ω-domain

Laplace Transform $\quad F(s) = \int_{-\infty}^{\infty} \exp(-st)\, f(t)\, dt \qquad$ s is $a+bi$ (complex number)

t-domain $\longleftrightarrow$ s-domain

Convolution $\quad (f * g)(t) = \int_{-\infty}^{\infty} g(t-\tau)\, f(\tau)\, d\tau$



General Integral Equation

$$y(x) = \int_a^x k(x,t)\, f(t)\, dt$$

limits can differ from $\pm\infty$

kernel function

$k(x,t) = g(x-t) \Rightarrow$ convolution

$k(x,t) = \exp(-xt) \Rightarrow$ Laplace

$K(x,t) = \exp(-ixt) \Rightarrow$ Fourier

$\vdots$

t - domain $\longleftrightarrow$ x - domain

Applications / Problem

- Measurement is in the x - domain $\Rightarrow$ quantity of interest is in t - domain
  - surface temperature evolution
  - size distribution function
  - optical measurements / remote sensing
  - FTIR
  - $\vdots$
- Inversion $\Rightarrow$ find $f(t)$ from discrete $y(x)$

Example Inversion Problem

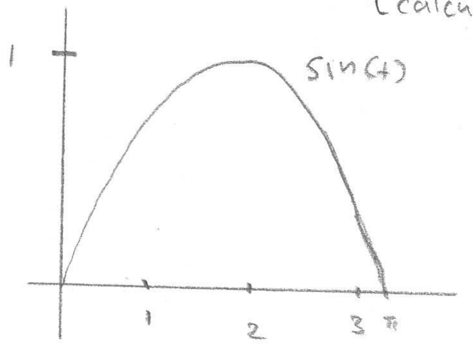
$$y(x) = \int_a^b k(x,t)\, u(t)\, dt \qquad \text{"Fredholm Integral Eq."}$$

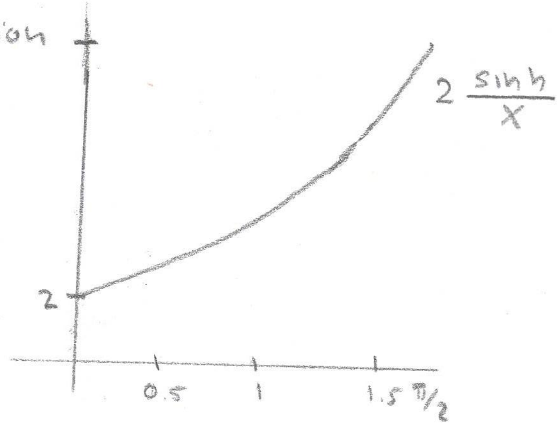
$$k(x,t) = \exp(x \cdot \cos(t)) \qquad u(t) = \sin(t) \qquad y(x) = \frac{2\sinh(x)}{x}$$

$a = 0,\ b = \pi \qquad t \in [0, \pi] \qquad$ and $\qquad x \in [0, \pi/2]$

test : $\displaystyle\int_0^\pi \exp(x\cos(t))\sin(t)\, dt = \frac{2\sinh}{x}$

→ verify using integration (calculus)





t - domain ⟷ x - domain

Discretized Integral Equation

$$\int_a^b k(x,t)u(t)\,dt = y(x)$$

$$A\vec{u} = \vec{y}$$

$\vec{u} = [u_1, \ldots, u_n]$

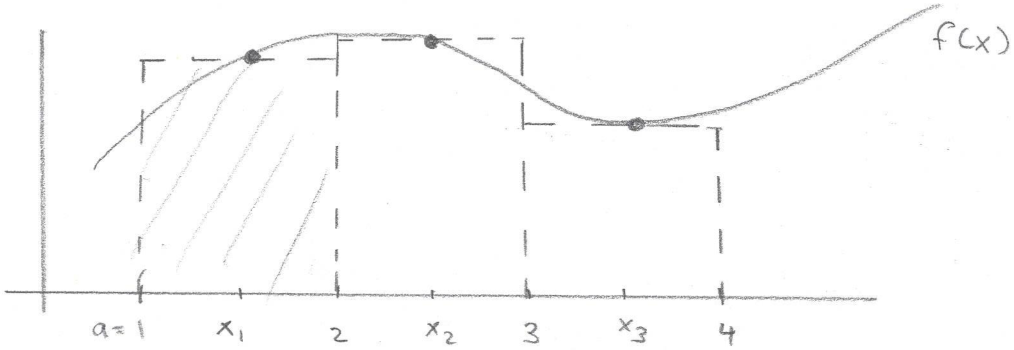
$\vec{y} = [y_1, \ldots, y_m]$

$A = \mathbb{R}^{m \times n}$

Question: how to find A ? (next page)

$$\sum_{i=1}^{n} w_n k(x_j, t_i) u(t_i) \approx y_j$$

Numerical Integration



$n = 3$: number of data points

$a = 1$, $b = 4$: limits of integration

$w = \frac{b-a}{n}$ bin width (here $w = 1$)

$$\int_a^b f(x)\,dx = \sum_{i=1}^{n} f(x_i)\,w$$

quadrature algorithm

$$x_i = (i - \tfrac{1}{2})\,w + a$$

$x_1 = (1 - 0.5)1 + 1 = 1.5$

$x_2 = (2 - 0.5)1 + 1 = 2.5$

$x_3 = (3 - 0.5)1 + 1 = 3.5$

$$\int_a^b k(x,t)\,u(t)\,dt = y(x) \qquad t \in [a,b] \qquad x \in [c,d]$$

u is discretized into n points $i = 1..n$ $\qquad w_n = \dfrac{b-a}{n}$

y is discretized into m points $j = 1..m$ $\qquad w_m = \dfrac{d-e}{m}$

$$y_1 = \int_a^b k(x,t)\,u(t)\,dt$$

$$y_1 = k(x_1,t_1)\,u(t_1)\,w_n + k(x_1,t_2)\,u(t_2)\,w_n + \dots + k(x_1,t_n)\,u(t_n)\,w_n$$

$$\vdots$$

$$y_m = k(x_m,t_1)\,u(t_1)\,w_n + k(x_m,t_2)\,u(t_2)\,w_n + \dots + k(x_m,t_n)\,u(t_n)\,w_n$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} k(x_1,t_1)\,w_n & \cdots & k(x_1,t_n)\,w_n \\ \vdots & & \\ k(x_m,t_1)\,w_n & \cdots & k(x_m,t_n)\,w_n \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}$$

$a_{j,i} = w_n\,k(x_j,t_i)$

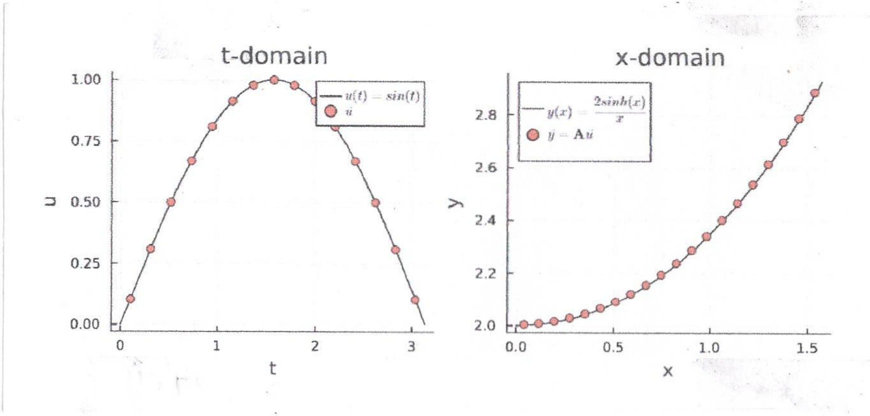
$A = \mathbb{R}^{m \times n}$

$$\underset{\mathbb{R}^m}{y} = A\,u \qquad u = \mathbb{R}^n$$

Example Discretization

→ n, m can be independently varied

→ overdetermined and underdetermined transforms



$n = 15$ $m = 20$

Example Inversion

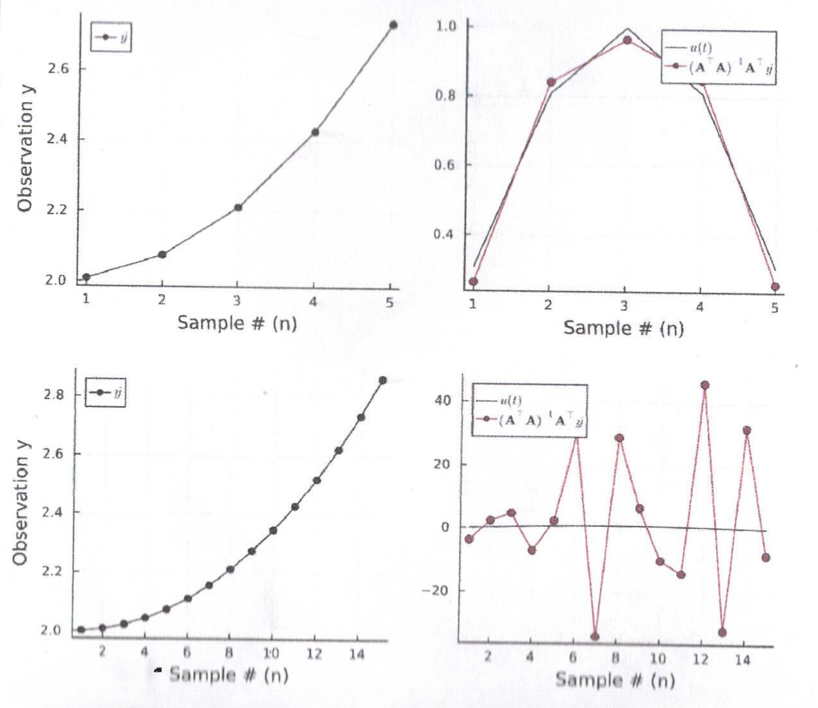
$y = A u(t)$

SSE: minimize $\|Au - y\|^2$

find $\vec{u}$ that is best

Observation in X-domain

prediction in t-domain



$u = (A^T A)^{-1} A^T y$

works well for small n

fails for large n

why?

Ill posed Problems

- Lack of a unique solution
- Sensitivity to initial conditions (noise) -> weather prediction
- nonexistence of a solution

$A = \begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix}$ det $A = 0$

$A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$ det $A = 0$

$\Big\}$
- underdetermined
- no unique solution

A^{-1} not defined

rank(A) = maximun number of linearly independent columns

here rank(A) = 1

Full rank = 2 (max possible)

A is "rank deficient"

Moore - Penrose Inverse

$$A^+ = (A^TA)^{-1}A^T$$

(1) for $m=n$ $A^+ = A^{-1}$ if A is full rank

2) $m>n$: • over determined system (e.g. linear regression case)
- more eq. than unkowns
- minimizes $\|Au - y\|^2$

(3) $n>m$: • under determined system
- fewer eq. than unkowns
- infinite solutions

- if A is not full rank A^+ must be computed using SVD
- use pinv(A) in most languages

Singular Value Decomposition

Full SVD

$$\underbrace{A}_{\mathbb{R}^{m\times n}} = \underbrace{U}_{\substack{\mathbb{R}^{m\times m}\\ \text{left singular vectors}}} \overbrace{\Sigma}^{\substack{\text{singular values}\\ \mathbb{R}^{m\times n}}} \underbrace{V^{T}}_{\substack{\mathbb{R}^{n\times n}\\ \text{right singular vectors}}}$$

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & & \sigma_p & 0 \end{bmatrix}$$

$\sigma_1 > \sigma_2 \dots > \sigma_p$

$p = \min\{m, n\}$

- can find using numerical methods
- $O(> n^2) \Rightarrow$ expensive
- use svd(A) from linear algebra

Moore - Penrose Inverse

$$A^+ = V \Sigma^+ U^T$$

$$\Sigma^+ = \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & & & \vdots \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_p} & 0 \end{bmatrix}$$

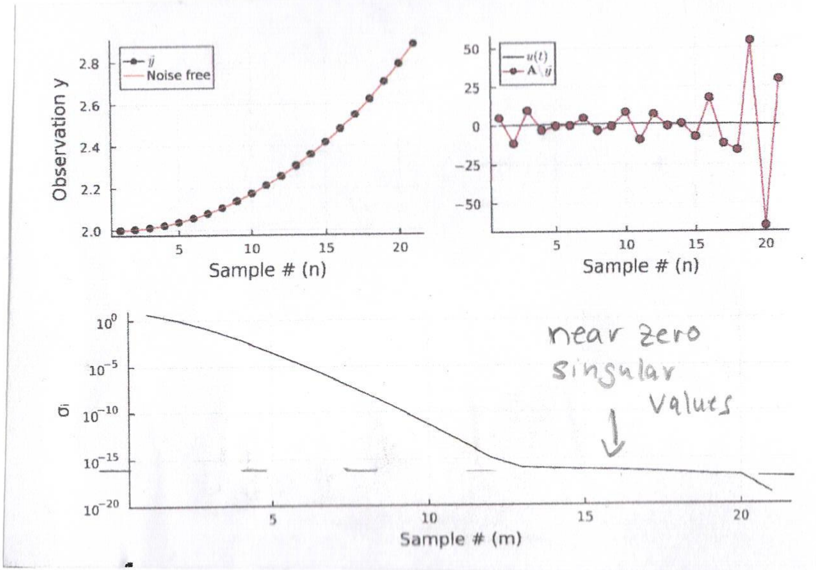
- numerical method to find A^+
- if $\sigma_i = 0$ then a zero is placed
- if A is rank deficient one or more $\sigma_i = 0$

Inversion Revisited

$\frac{1}{10^{-15}} = 10^{15} \Rightarrow$ large entries in Σ^+

"infected by numerical or measurement noise"

x - domain　　　　t - domain



near zero singular values

$$u = \overbrace{(A^T A)^{-1} A^T}\, y$$

$$u = A^+ y$$

log scale

machine precision for 64 bit floating point $\approx 10^{-16}$

no new information

Regularized Inverse (Noise Filtering)

filter matrix

initial guess

$$u_\lambda = \text{minimize}\left\{ \underbrace{\|Au - y\|^2}_{\text{ordinary regression}} + \lambda^2 \|L(u - u_0)\|^2 \right\}$$

regularization parameter

$$u_\lambda = (A^T A + \lambda^2 L^T L)^{-1} (A^T y + \lambda^2 L^T L u_0)$$

$$u_0 = A^+ y \quad !$$

$$\lim_{\lambda \to \infty} u_\lambda = u_0$$

Starting point for $L = I$

identity matrix

$u_0 = 0$

(no initial guess)

then

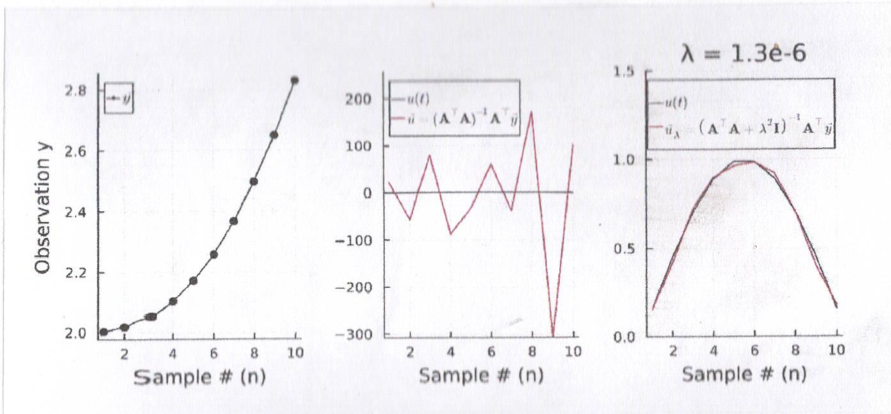
$$u_\lambda = (A^T A + \lambda^2 I)^{-1} A^T y \quad \text{"Ridge Regression"}$$

Inversion Revisited

- need to try range of λ
- $\lambda = 0 \Rightarrow$ bad solution
- λ large $\Rightarrow$ approach zero $\Rightarrow$ bad solution
- a λ exists to find good inversion

$\Rightarrow$ how to find λ ?

x-domain t-domain



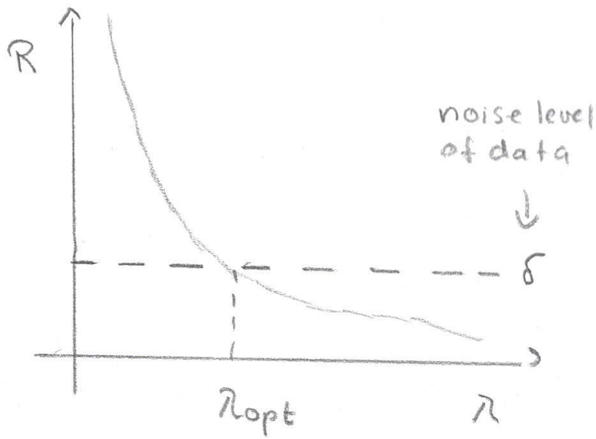
$\lambda = 0$

$$u_\lambda = (A^TA + \lambda^2 I)^{-1} A^T y$$

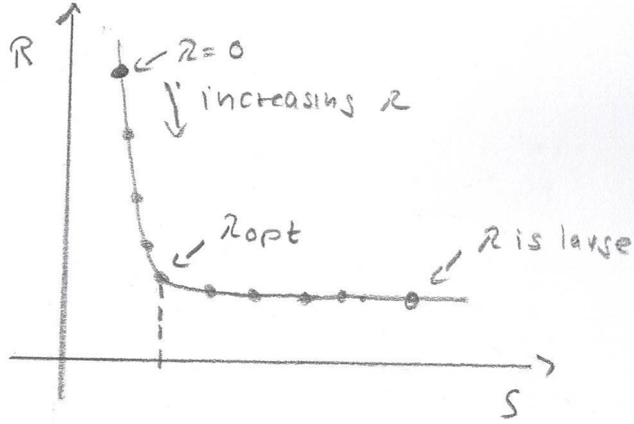
Methods to find λ

$$u_\lambda = \text{minimize}\left\{ \underbrace{\|Au - y\|^2}_{R = \text{Residual Norm}} + \lambda^2 \underbrace{\|L(u-u_0)\|^2}_{S = \text{Solution Norm}} \right\}$$

Morozev



L-curve



- λ_{opt} is "corner" of L-curve
- find visually or algorithm