簡單線性迴歸 最小平方方法

課程：量物實驗

教師：楊仲準

迴歸分析 (regression analysis)

- 迴歸分析： 以數學和統計方法來確認一組變數中的系統性部分，並依此解釋過去的現象和預測未來。

- 估計迴歸係數最常用的方法之一就是**普通最小平方(ordinary least squares)**，又簡稱為**最小平方法**。

簡單線性迴歸模型

- 簡單線性迴歸模型：利用一個線性模型來捕捉 $\{(X_i,Y_i),i=1,..,n\}$ 這組雙變量隨機變數中 Y_i 的系統性部分 $g(X_i)$。
- $g(X)= a + bX$，其中 a，b 為未知參數，需要估計。
- 可以將原始 Y 表示為

$$\boldsymbol{Y = a + bX + U},$$

其中 U 代表不能由 $a + bX$ 所描述的 Y 行為，亦即 Y 與線性模型之間的誤差。

簡單線性迴歸模型

$Y = a + bX + U$

- 迴歸模型中的變數 Y 稱作應變數 (dependent variable 或 regressand)
- 變數 X 稱作解釋變數 (explanatory variable 或 regressor)。
- 參數 a 和 b 稱作迴歸係數 (regression coefficient)。
  - a: 截距項，
  - b: 斜率。

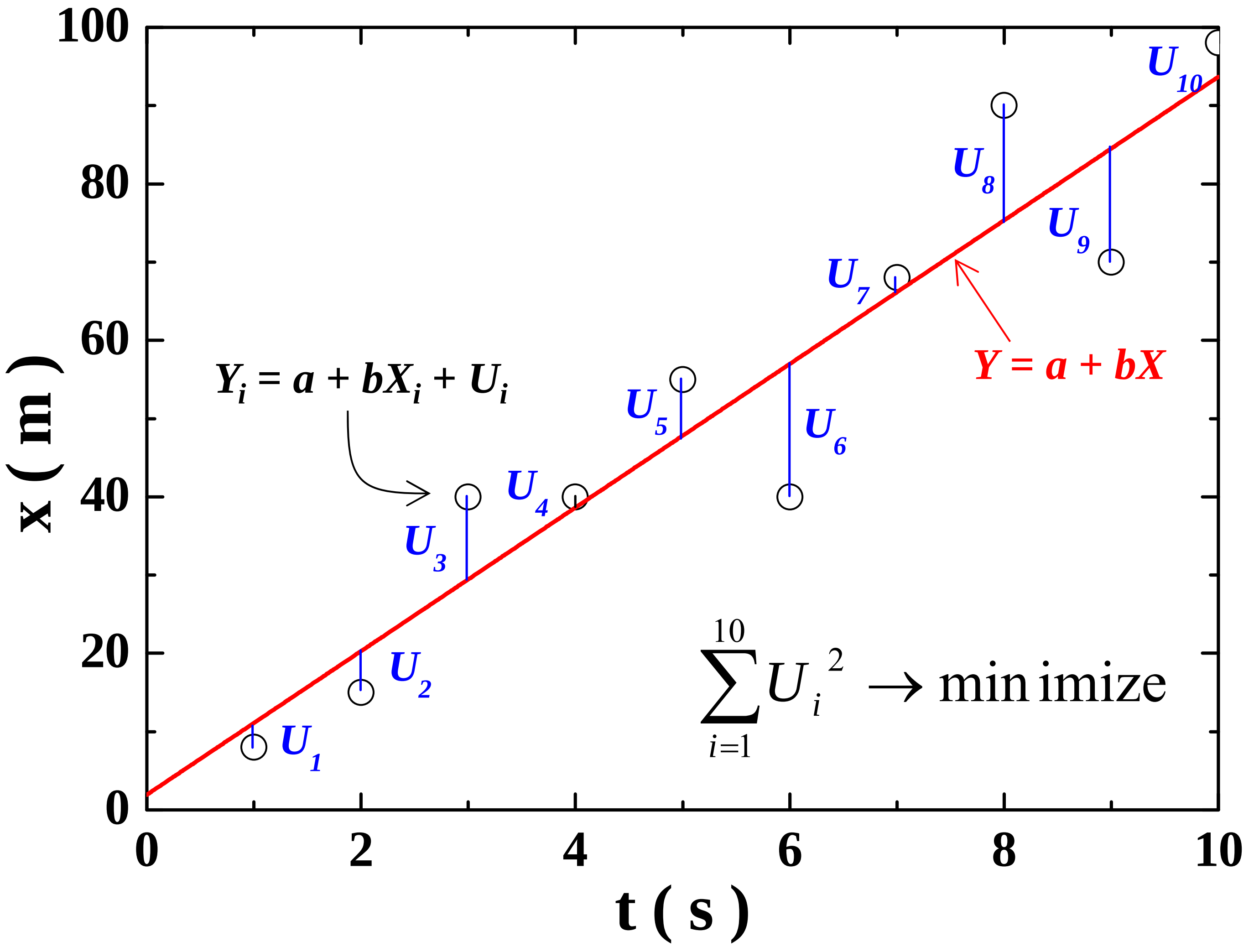
最小平方法原理

- 找a和b使模型誤差 U_i 的平方和極小。採用誤差平方和是為了避免正負誤差之間互相抵銷。
- 目標函數如下：

$$Q(a,b)=\frac{1}{n}\sum_{i=1}^{n}(Y_i-a-bX_i)^2 \quad =\frac{1}{n}\sum_{i=1}^{n}U_i^2.$$

- 最小平方法所找的就是使誤差平方和（或其平均）最小的那條直線。

最小平方法原理



100
80
60
40
20
0
0
2
4
6
8
10
x (m)
t (s)
$Y_i = a + bX_i + U_i$
U_1
U_2
U_3
U_4
U_5
U_6
U_7
U_8
U_9
U_{10}
$Y = a + bX$
$\sum_{i=1}^{10} U_i^2 \rightarrow \text{minimize}$

最小平方法原理

- 為使目標函數之值最小，必須解出以下的一階條件 (first order condition)。

$$\frac{\partial}{\partial a}\mathrm{Q}(a,b) = -2\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{Y}_{\mathrm{i}} - a - b\mathrm{X}_{\mathrm{i}}) = 0,$$

$$\frac{\partial}{\partial b}\mathrm{Q}(a,b) = -2\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{Y}_{\mathrm{i}} - a - b\mathrm{X}_{\mathrm{i}})\mathrm{X}_{\mathrm{i}} = 0.$$

- 這兩個一階條件又稱作**標準方程式 (normal equations)**。

最小平方法原理

- 可從標準方程式中求出a和b的解，稱作最小平方估計式(ordinary least squares estimator，簡稱OLS estimator)，一般以

$$b=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sum_{i=1}^{n}(X_i-\overline{X})^2},\qquad a=\overline{Y}-b\overline{X}.$$

其中

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i\qquad \overline{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$$

- 若 X_i 為常數，$X_i=\overline{X}$ 則 $\hat{\beta}_n$ 根本無法計算。因此需要一個「認定條件」：
  - X_i , $i=1,2,\ldots,n$ 之值不為常數

最小平方法：迴歸直線

- 將最小平方估計式a與b代入設定的線性模型就可得到一條截距為a，斜率為b的直線，稱作**估計的迴歸線(estimated regression line)**。

$$\boxed{Y = a + bX}$$

- 斜率係數估計式b衡量X的邊際效果：當X變動一單位時，估計的迴歸線會預測應變數Y將變動b個單位。
- 截距係數a則表示當X為0時，估計的迴歸線所預測的應變數Y。
- 將樣本中的變數X_i代入估計的迴歸線，即可求得估計的應變數。

最小平方法：迴歸直線(整理)

$y = (a \pm \sigma_a) + (b \pm \sigma_b)x$

(完美配適) $1 > R^2 > 0$ (無對應關係)

1. $$\begin{cases} \bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i \\ \bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i \end{cases}$$

平均值

2. $$\begin{cases} b = \dfrac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{\sum_{i=1}^{n}(x-\bar{x})^2} \\ a = \bar{y} - b\bar{x} \end{cases}$$

迴歸係數

3. $$s = \sqrt{\dfrac{\sum_{i=1}^{n}[y_i - (a + bx_i)]^2}{n-2}}$$

標準差

4. $$\begin{cases} \sigma_b = \dfrac{s}{\sqrt{\sum_{i=1}^{n}(x-\bar{x})^2}} \\ \sigma_a = s\sqrt{\left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{\sum_{i=1}^{n}(x-\bar{x})^2}\right)} \end{cases}$$

迴歸係數誤差

5. $$R = \dfrac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y-\bar{y})^2}}$$

判定係數

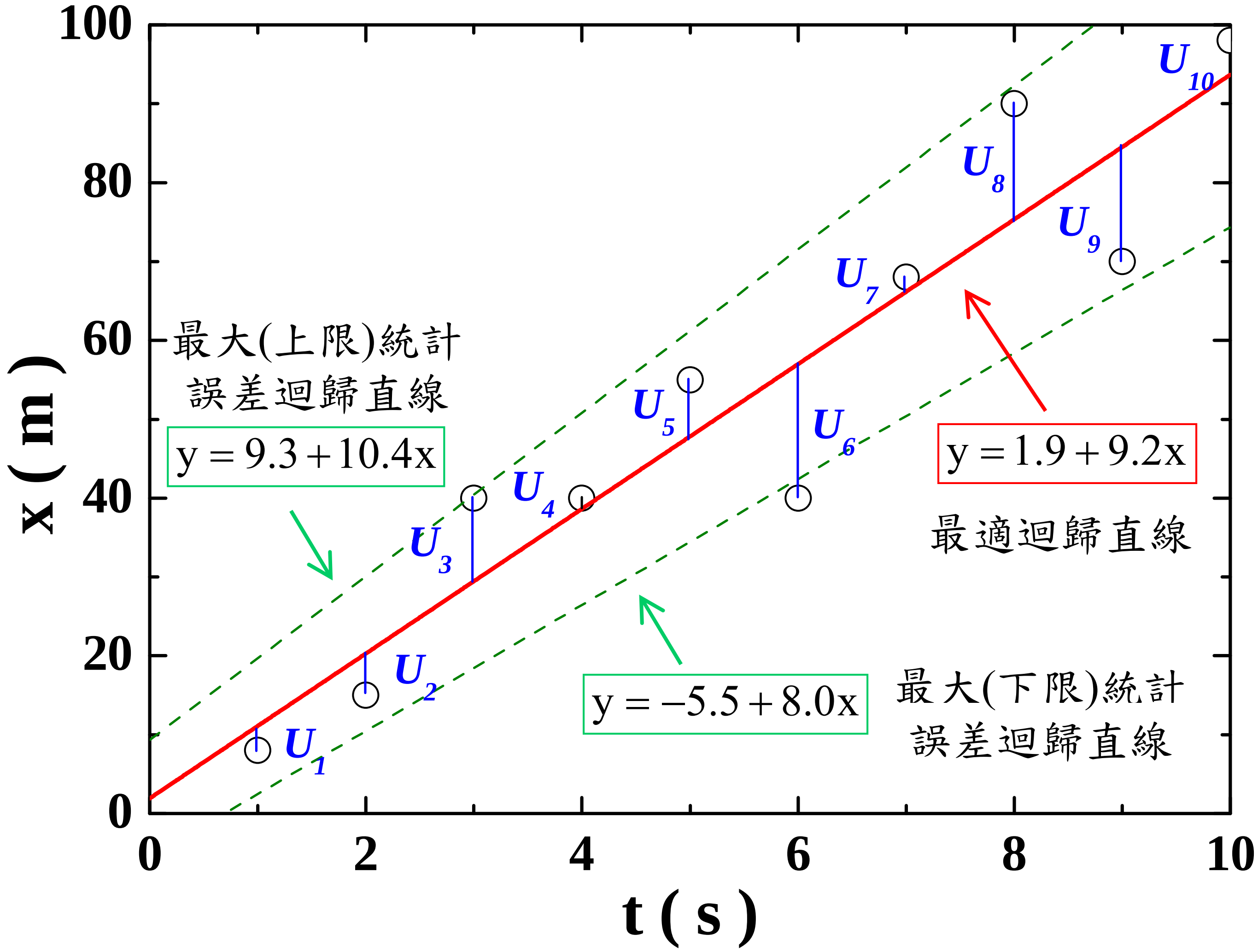
迴歸直線：(等速度運動)

t (s)	x (m)	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$[y_i-(a+bx_i)]^2$
1.0	8.0	-4.5	-44.4	199.8	20.25	1971.36	9.67
2.0	15.0	-3.5	-37.4	130.9	12.25	1398.76	27.93
3.0	40.0	-2.5	-12.4	31	6.25	153.76	111.08
4.0	40.0	-1.5	-12.4	18.6	2.25	153.76	1.86
5.0	55.0	-0.5	2.6	-1.3	0.25	6.76	51.67
6.0	40.0	0.5	-12.4	-6.2	0.25	153.76	288.59
7.0	68.0	1.5	15.6	23.4	2.25	243.36	3.37
8.0	90.0	2.5	37.6	94	6.25	1413.76	214.93
9.0	70.0	3.5	17.6	61.6	12.25	309.76	210.69
10.0	98.0	4.5	45.6	205.2	20.25	2079.36	18.57
$\bar{x}=$ **5.5**	$\bar{y}=$ **52.4**			$\sum=$ **757**	$\sum=$ **82.5**	$\sum=$ **7884.4**	$\sum=$ **938.35**

$$y=(1.9\pm 7.4)+(9.2\pm 1.2)x$$

$$R^2=0.88$$

迴歸直線：(等速度運動)



100
80
60
40
20
0
0
2
4
6
8
10
x (m)
t (s)
U1
U2
U3
U4
U5
U6
U7
U8
U9
U10
最大(上限)統計
誤差迴歸直線
y = 9.3 + 10.4x
y = 1.9 + 9.2x
最適迴歸直線
y = −5.5 + 8.0x
最大(下限)統計
誤差迴歸直線

最小平方法：EXCEL操作 (常用指令)

- Σ:總和，其指令形式為sum(起始資料:終止資料)
- SQRT(資料)：開根號
- ˆn, n為數字:代表次方，如5^3=5^3
- $:鎖定行或是列，如
  - $B10(鎖定B不隨滑鼠拖曳變化)
  - B$10 (鎖定10不隨滑鼠拖曳變化)
  - B10 (鎖定B10不隨滑鼠拖曳變化)
- Pi()=π=3.1415926
- Average(起始資料:終止資料)：求平均

最小平方法：EXCEL操作

	A	B	C	D	E	F	G	H
1	t	x	x-xbar	y-ybar	(x-xbar)(y-ybar)	$(x\text{-}xbar)^2$	$(y\text{-}ybar)^2$	$[y_i-(a+bx_i)]^2$
2	1.0	8.0	-4.5	-44.4	199.8	20.25	1971.36	9.67
3	2.0	15.0	-3.5	-37.4	130.9	12.25	1398.76	27.93
4	3.0	40.0	-2.5	-12.4	31	6.25	153.76	111.08
5	4.0	40.0	-1.5	-12.4	18.6	2.25	153.76	1.86
6	5.0	55.0	-0.5	2.6	-1.3	0.25	6.76	51.67
7	6.0	40.0	0.5	-12.4	-6.2	0.25	153.76	288.59
8	7.0	68.0	1.5	15.6	23.4	2.25	243.36	3.37
9	8.0	90.0	2.5	37.6	94	6.25	1413.76	214.93
10	9.0	70.0	3.5	17.6	61.6	12.25	309.76	210.69
11	10.0	98.0	4.5	45.6	205.2	20.25	2079.36	18.57
12	5.5	52.4			757	82.5	7884.4	938.35
13	b	9.2	σ_b	1.2				
14	a	1.9	σ_a	7.4				
15	s	10.8	R^2	0.88				

2
=B2-B12
=C2*D2
=C2^2
=D2^2
=(B2-(B14+B13*A2))^2
=A2-A12

這一排設定好了以後全選然後用滑鼠由2向下拖曳到11

1
=AVERAGE(A2:A11)
=AVERAGE(A2:A11)

3
=SUM(H2:H11)
=SUM(G2:G11)
=SUM(F2:F11)
=SUM(E2:E11)

4
=SQRT(H12/(10-2))
=B12-B13*A12
=E12/F12
=(E12^2)/(F12*G12)
=B15*SQRT((1/10)+(A12^2)/F12)
=B15/SQRT(F12)

迴歸直線：(範例-等速度運動)

- 已知平均速度的定義為:

$$\bar{v} \equiv \frac{x_f - x_i}{t_f - t_i} = \frac{\Delta x}{\Delta t}$$

- 代表x-t圖上的斜率
- 又已知$x=x_0+vt$
- 上圖中，迴歸所得之方程式為y=(1.9±7.4)+(9.2±1.2)x

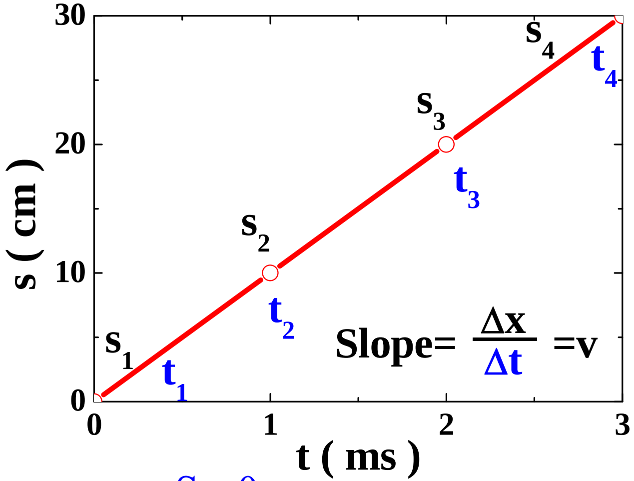
其中應變數y代表x(m)，解釋變數x代表t(s)。所以此一迴歸公式所代表的物理式子為：

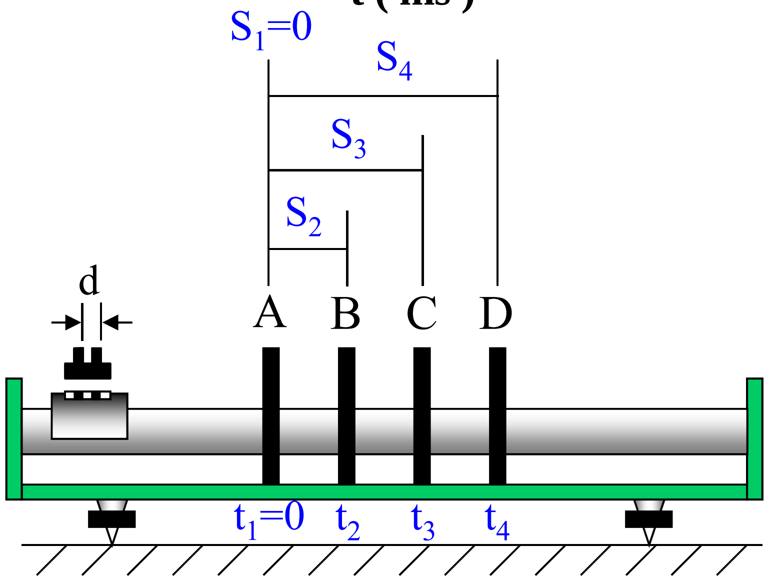
$$x=(1.9\pm7.4)+(9.2\pm1.2)t$$

- 因此，本次實驗所得到的起始位置
  - $x_0=(1.9\pm7.4)$ m
  - $v=(9.2\pm1.2)$ m/s

迴歸直線：(範例-等速度運動)

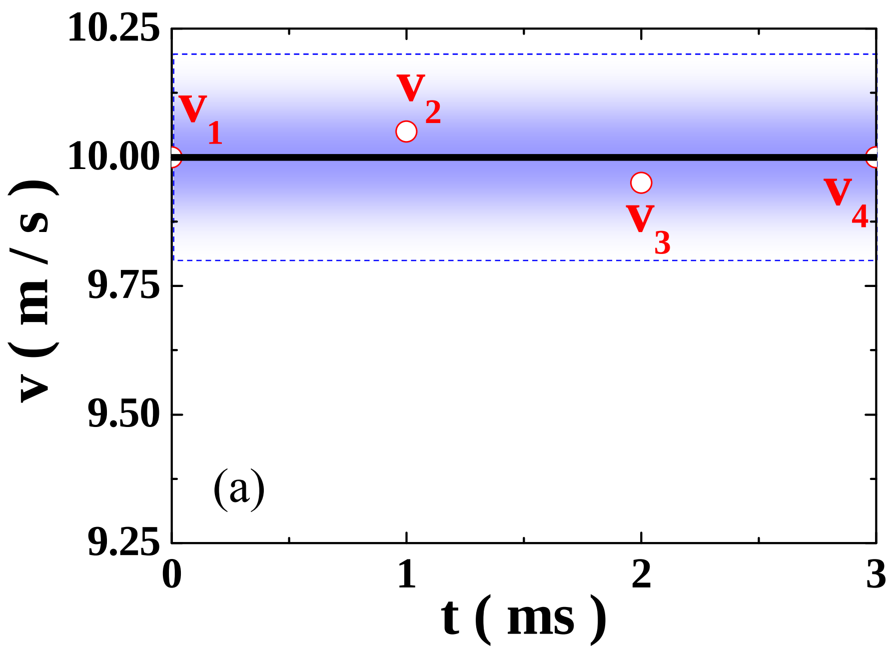
- 電腦上所得到的數值意義
  - s：各光電門與光電門A的位置差
  - Δt：滑塊上檔版的兩個前沿通過某一個光電門的時間
  - d：滑塊上檔版的兩個前沿間距
  - v：定義為d/Δt
  - t：由光電門A到通過另一個光電門所需的時間
- 因此v趨近於滑塊通過某光電門的瞬時速度。
- 由於等速度運動期間，瞬時速度等於平均速度。因此，對s-t圖做線性迴歸之斜率應該接近個別光電門所的到之v值。

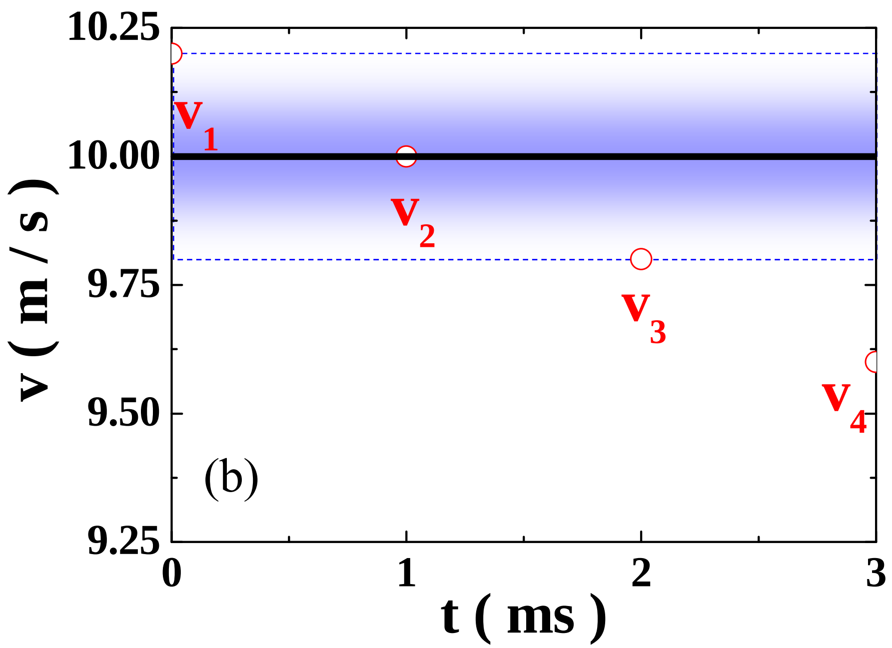




迴歸直線：(範例-等速度運動)

- 將迴歸直線所得到之v值做成v-t圖，並將四個光電門所得到之瞬時速度點畫上，應可發現圖形會形成類似圖(a)的樣子:瞬時速度接近平均速度，其起伏均落在誤差範圍內。
- 如果外力(如摩擦力)作用於滑塊上(造成減速)則可以發現類似圖(b)之現象。數據點形成遞減的狀態，甚至超出估計的誤差。同時R^2值(判定係數)也會出現與1相差較多的狀況。



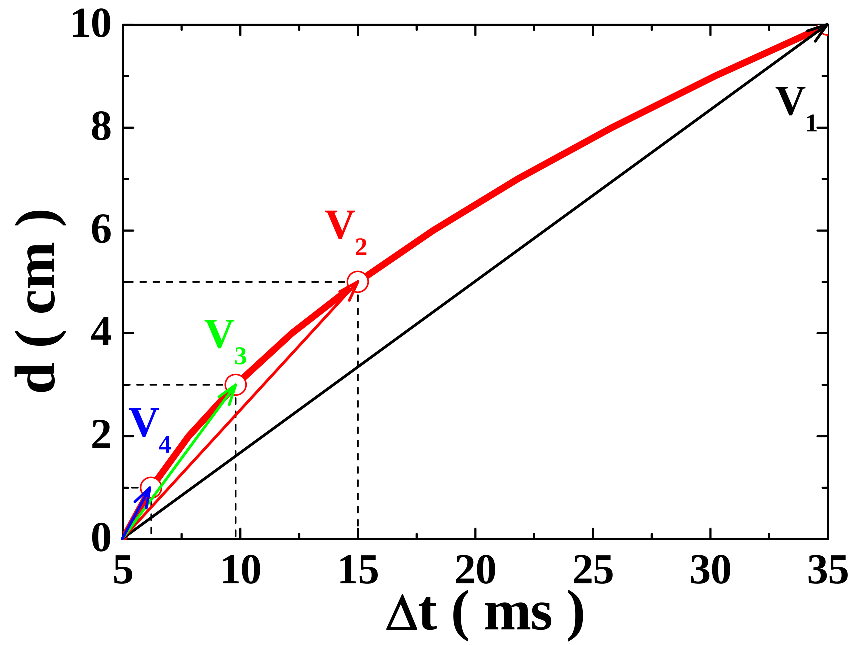


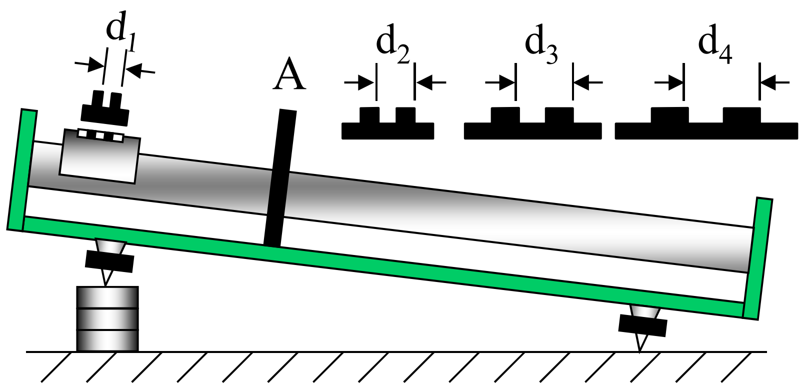
迴歸直線：(範例-瞬時速度)

- 電腦上所得到的數值意義
  - d：滑塊上檔版的兩個前沿間距
  - Δt：滑塊上檔版的兩個前沿通過某一個光電門的時間
  - v：定義為d/Δt
- 瞬時速度的定義為：

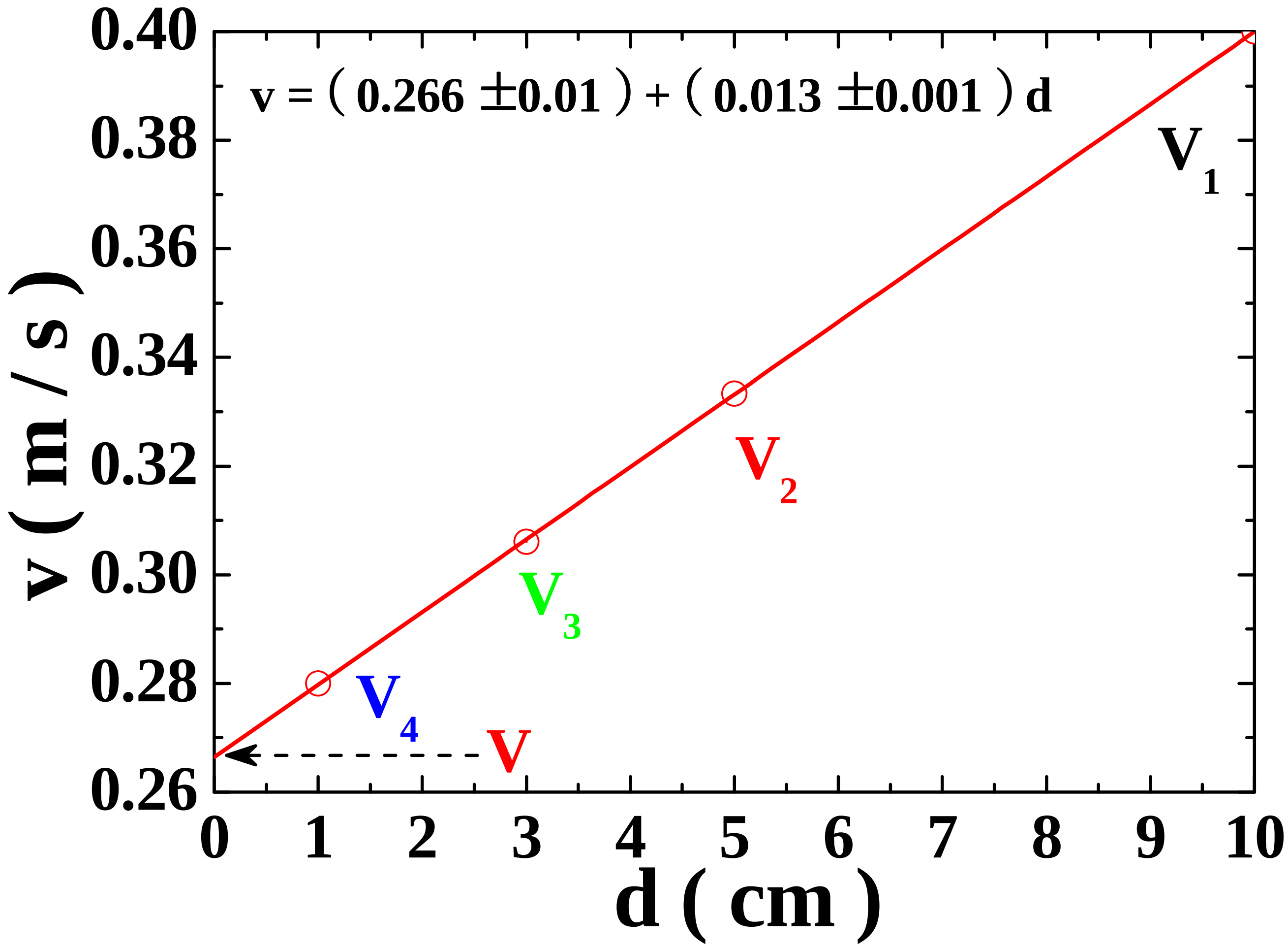
$$v \equiv \lim_{\Delta t \to 0} \frac{\Delta x}{\Delta t} = \frac{dx}{dt}$$

- 因此當d漸漸縮小時，v趨近於滑塊通過某光電門的瞬時速度





迴歸直線：(範例-瞬時速度)



迴歸直線的截距代表當d→0時的速度，也就是瞬時速度

簡單線性迴歸：最小平方法
結束