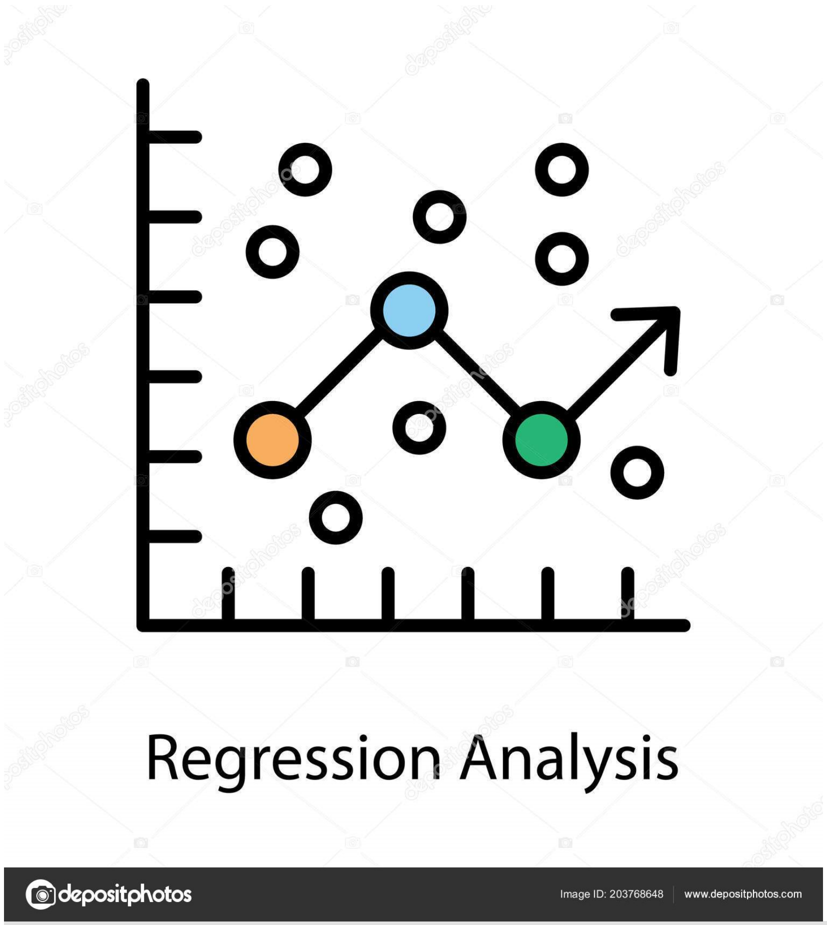
# Engineering Statistics




# Regression Analysis

**Dr. Vvn Weian Chao (趙韋安)**
https://ce.nctu.edu.tw/member/teachers/23
Department of Civil Engineering, National Yang Ming Chiao Tung University, Taiwan

# Purpose

# 單變數
# 進入
# 多變數

多變數關係

關係強度

# R: Useful function

cor().-相關係數

lm().-迴歸分析

anova().-變異數分析

lowess().-平滑曲線

# Outline

-Scatter plots
-Correlation
-Fitting a Line to Bivariate Data
-Nonlinear Relationships
-Using More than One Predictor

# Scatter plots

應變數(Y): Ocular Surface Area, 視覺表面積[$cm^2$]

自變數(X): 眼睛之間的寬度 [cm]

資料個數: 30筆, data.xlsx

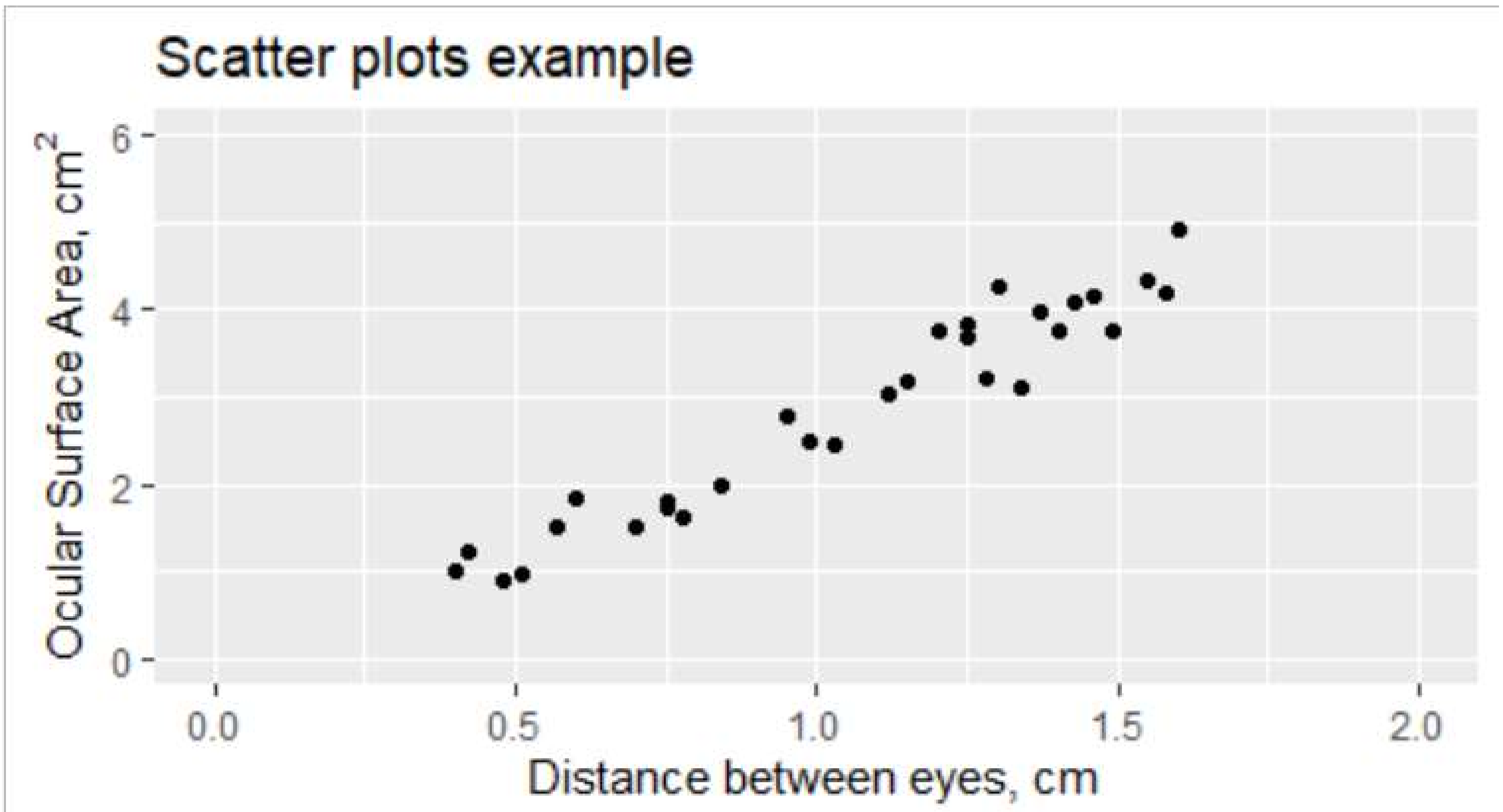
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Obs: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| *x*: | .40 | .42 | .48 | .51 | .57 | .60 | .70 | .75 | .75 | .78 |
| *y*: | 1.02 | 1.21 | .88 | .98 | 1.52 | 1.83 | 1.50 | 1.80 | 1.74 | 1.63 |
| Obs: | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| *x*: | .84 | .95 | .99 | 1.03 | 1.12 | 1.15 | 1.20 | 1.25 | 1.25 | 1.28 |
| *y*: | 2.00 | 2.80 | 2.48 | 2.47 | 3.05 | 3.18 | 3.76 | 3.68 | 3.82 | 3.21 |
| Obs: | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| *x*: | 1.30 | 1.34 | 1.37 | 1.40 | 1.43 | 1.46 | 1.49 | 1.55 | 1.58 | 1.60 |
| *y*: | 4.27 | 3.12 | 3.99 | 3.75 | 4.10 | 4.18 | 3.77 | 4.34 | 4.21 | 4.92 |

# Scatter plots

X增加, Y亦增加

線性關係?

是否通過原點? (**符合物理意義**)



# Correlation

Pearson’s Sample Correlation Coefficient

Properties & Interpretation of r

Correlation & Causation

https://zh.wikipedia.org/wiki/%E5%8D%A1%E5%B0%94%C2%B7%E7%9A%AE%E5%B0%94%E9%80%8A

## 卡爾·皮爾森 [編輯]

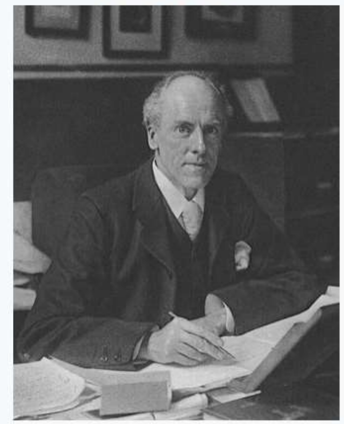
維基百科，自由的百科全書

**卡爾·皮爾森**（Karl Pearson，1857年3月27日 - 1936年4月27日），英國數學家和自由思想家。

目錄 [隱藏]



**卡爾·皮爾森**
**Karl Pearson**



| | |
|---|---|
| 出生 | Carl Pearson<br>1857年3月27日<br>英國倫敦伊斯靈頓 |
| 逝世 | 1936年4月27日（79歲）<br>英國薩里郡多爾金 |
| 國籍 | 英國 |
| 母校 | 劍橋大學 |

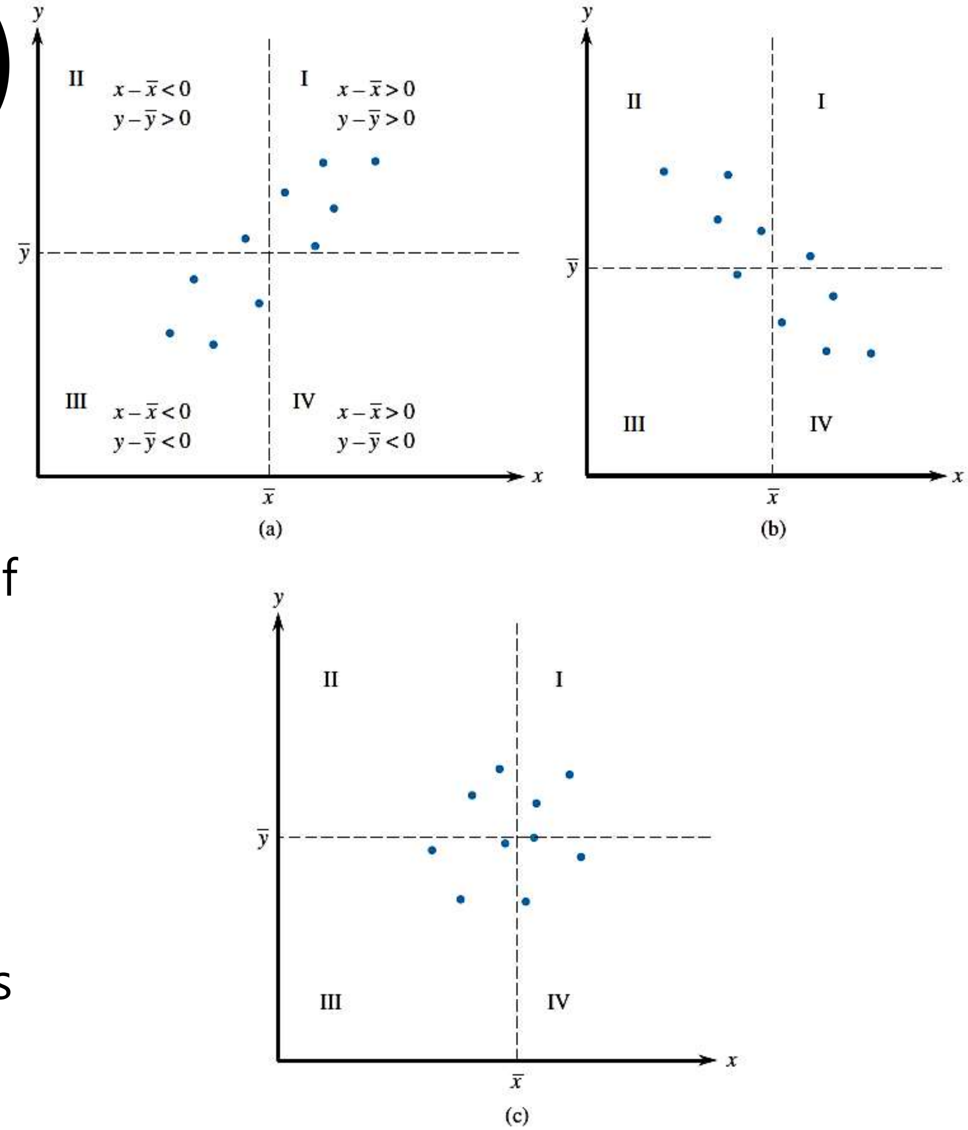
### 生平 [編輯]

1857年出生於英國倫敦；1879年畢業於劍橋大學，獲數學學士學位；[1]後往德國海德堡大學進修德語及人文學科；後去林肯法學院學習法律獲大律師資格；數年後於劍橋大學獲數學哲學博士學位；1884年～1911年任倫敦大學應用數學和力學的教授，1911年～1933年任高爾頓實驗室主任，又任應用統計系教授。

1896年選為英國皇家學會會員，他還是愛爾堡皇家

# Pearson Correlation Coefficient (r)

- ✓ Multiply each x deviation by the corresponding y deviation to obtain products of deviations of the form (x-x_avg)(y-y_avg).
- ✓ **Fig.a**, because almost all points lie in **regions I and III**, almost all products of deviations are positive. Thus the sum of products will be **a large positive number**.
- ✓ **Fig.b** exhibits a strong **negative** relationship.
- ✓ **In Fig.c**, positive and negative products of deviations tend to counteract on another, giving a value of the sum that is **close to zero**.



# Pearson Correlation Coefficient (r)

Pearson's sample correlation **r** is given by

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

Computing formulas for the three summation quantities are

平方和

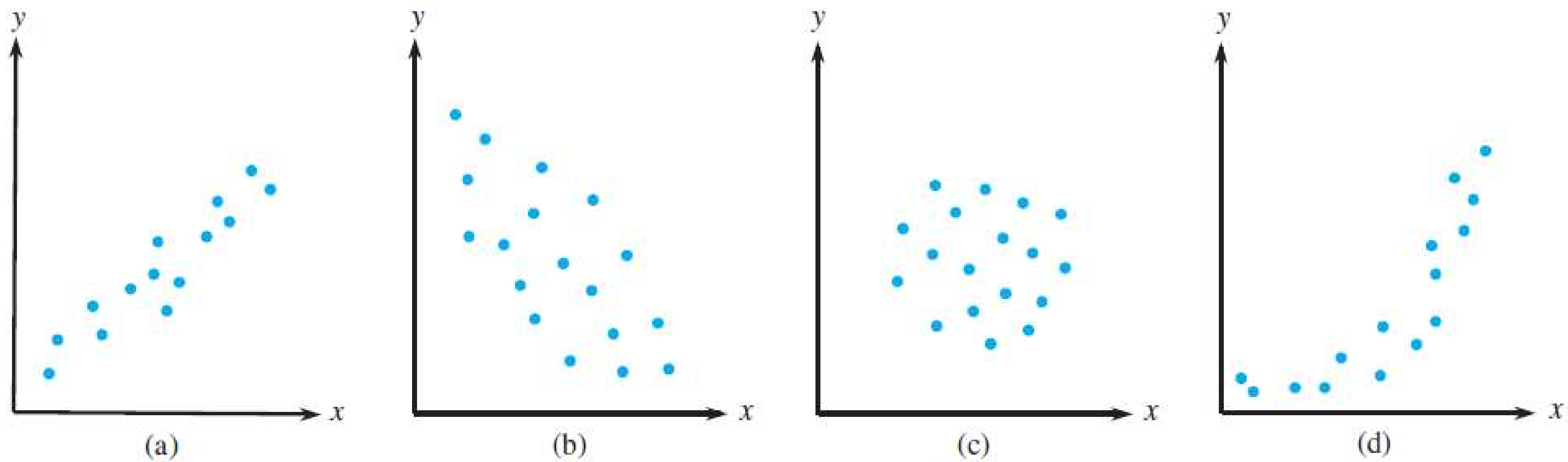
和平方

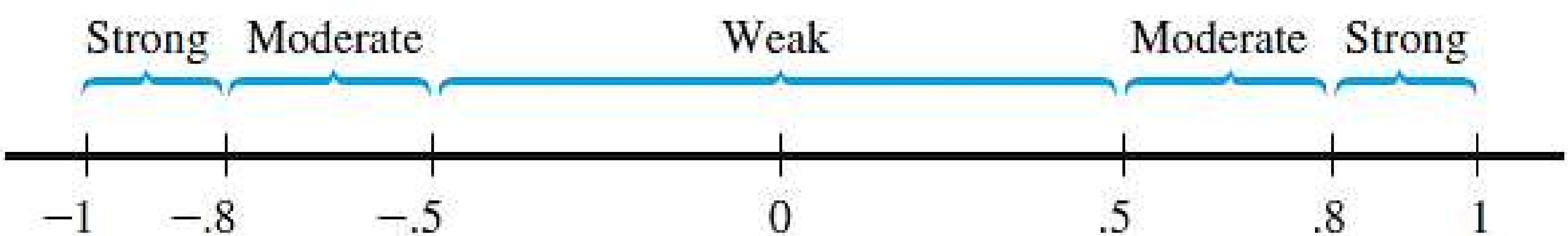
$$S_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$$S_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$$

$$S_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

# Properties & Interpretation of *r*



# R: Correlation Coefficient

# cor().

# R語言使用希臘字母與上下標符號 expression

-Use the expression(): ^上標 [ ]下標 ~空格 *連接符號
 expression(“r”^2~”= 0.123”)
-> $r^2$ = 0.123

# TRY it in R

# R: Correlation Coefficient

R_regression_a.R

# Correlation & Causation

實際上，雖觀察到兩變數具有極高相關係數，並非真正代表兩個變數之間確實存在因果關係

因為，有可能兩個變數同時與第三個變數存在強烈的關係

小孩牙齒數量與說話能力有正相關。但是，實際上牙齒數量與說話能力皆與年齡有明確關係，當分析時固定年齡，則兩者變數之間關係會變小

# Fitting a Line to Bivariate Data

Fitting a Straight Line

Assessing the Fit of the Least Squares Line

Standard Deviation about the LS Line

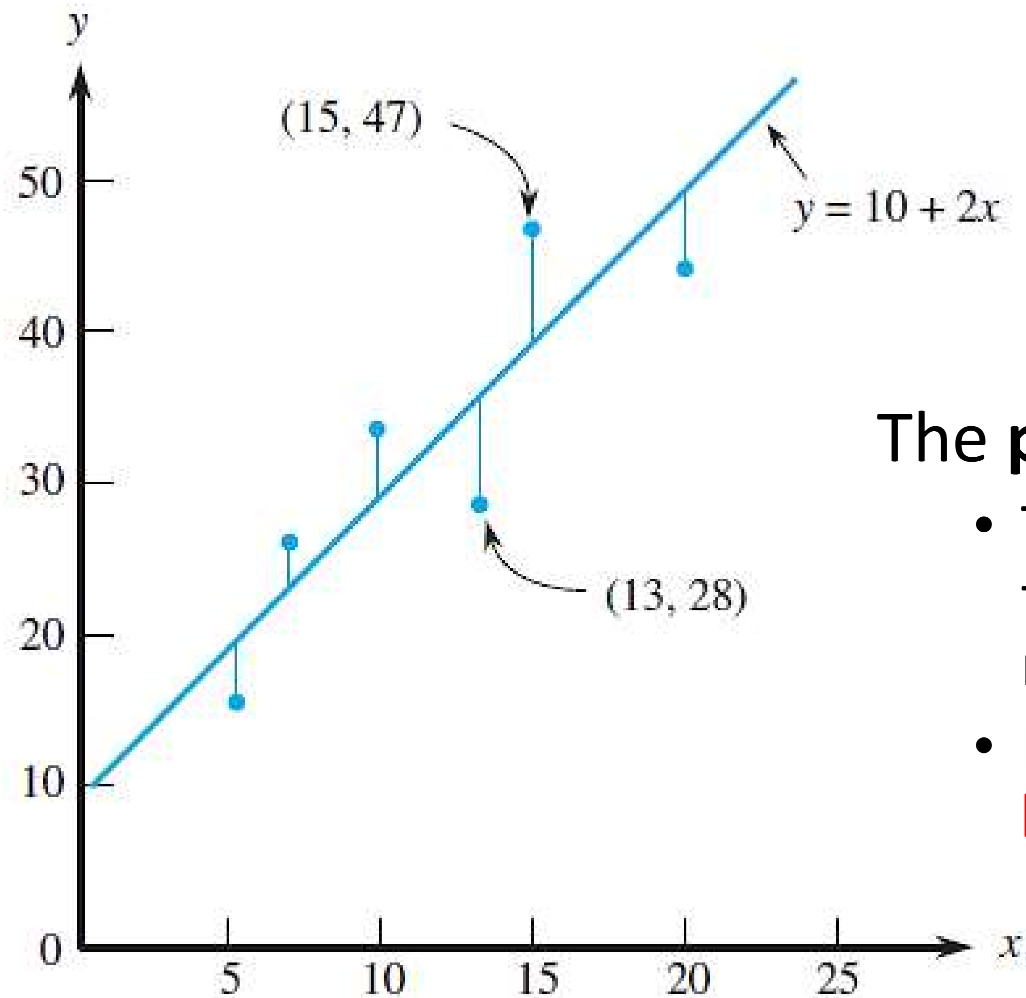
Plotting the Residuals

Resistant Lines

$$\boxed{Y = \text{SLOPE} \times X + \text{INTERCEPT}}$$

# Fitting a Straight Line

$$\sum[y_i-(a+bx_i)]^2=[y_1-(a+bx_1)]^2+\ldots+[y_n-(a+bx_n)]^2$$




$$\boxed{y=a+bx}$$

The **principle of least squares:**

- The line that gives the best fit to the data is the one that minimizes the sum.
- It is called **the least squares line** or **sample regression line.**

# Fitting a Straight Line

$$\sum[y_i-(a+bx_i)]^2=[y_1-(a+bx_1)]^2+\ldots+[y_n-(a+bx_n)]^2$$

$$\boxed{\begin{cases}\dfrac{d}{da}\sum[y_i-(a+bx_i)]^2\\[2ex]\dfrac{d}{db}\sum[y_i-(a+bx_i)]^2\end{cases}\Rightarrow\begin{cases}na+(\sum x_i)b=\sum y_i\\(\sum x_i)a+(\sum x_i^2)b=\sum x_iy_i\end{cases}}$$

(1)解二元一次方程組：$\begin{cases}a_1x+b_1y=c_1\cdots(1)\\a_2x+b_2y=c_2\cdots(2)\end{cases}$，其中$x,y$是未知數，

我們使用代入消去法解之

$(1)\times b_2-(2)\times b_1\Rightarrow(a_1b_2-a_2b_1)x=(c_1b_2-c_2b_1)$

$(1)\times a_2-(2)\times a_1\Rightarrow(a_2b_1-a_1b_2)y=(c_1a_2-c_2a_1)$

$\Rightarrow$可得$\begin{cases}\Delta\cdot x=\Delta_x\\\Delta\cdot y=\Delta_y\end{cases}$，其中$\Delta=\begin{vmatrix}a_1&b_1\\a_2&b_2\end{vmatrix}$，$\Delta_x=\begin{vmatrix}c_1&b_1\\c_2&b_2\end{vmatrix}$，$\Delta_y=\begin{vmatrix}a_1&c_1\\a_2&c_2\end{vmatrix}$。

當$\Delta\neq0$時，方程組恰有一解$(x,y)=(\dfrac{\Delta_x}{\Delta},\dfrac{\Delta_y}{\Delta})$ **[兩直線交於一點]**

當$\Delta=\Delta_x=\Delta_y=0$，方程組有無限多解。**[兩直線重合]**

當$\Delta=0$，而$\Delta_x$、$\Delta_y$有一不爲 0 時，方程組無解。**[兩直線平行]**

# Fitting a Straight Line

- The slope *b* of the least squares line is:

$$b = \frac{\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)/n}{\sum x_i^2 - \left(\sum x_i\right)^2/n} = \frac{S_{xy}}{S_{xx}}$$

- The vertical intercept *a* of the least squares line is

$$a = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

$$a = \bar{y} - b\bar{x}$$

# Fitting a Straight Line

$$b = \frac{S_{xy}}{S_{xx}}; r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

If r/b, then

$$\frac{r}{b} = \frac{S_{xx}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}}$$

$$\boldsymbol{b = r\left(\frac{s_y}{s_x}\right)}$$

- 相關係數r控制斜率正負
- x,y變數的標準差及r控制斜率大小

# Assessing the Fit of LS Line

- Predicting the y values: $\hat{y}_i = a + bx_i$
- Residual sum of squares (殘差平方和), SSResid, SSE:

$$SSResid = SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Regression of sum of squares (迴歸平方和), SSR:

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- Total sum of squares (總平方和), SSTo:

$$SSTo = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

# Assessing the Fit of LS Line

$$SSTo = SSE + SSR$$

- 迴歸關係式無法解釋的資料比例: $SSE/_{SSTo}$
- Coefficient of Determination (決定係數): 可度量迴歸線對於資料的擬合程度

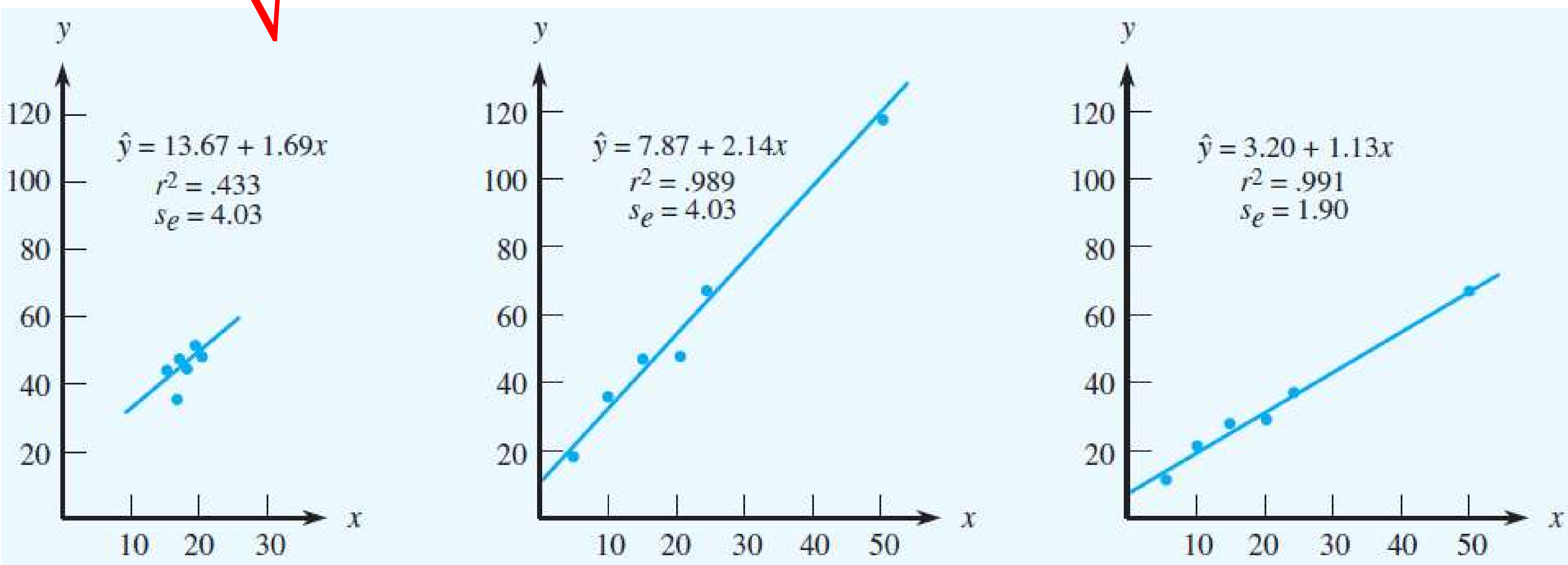
$$r^2 = 1 - \frac{SSE}{SSTo}$$

# Standard Deviation about the LS Line

$$b = r\left(\frac{s_y}{s_x}\right)$$

$$S_e = \sqrt{\frac{SSE}{n-2}}$$

度量觀測值在迴歸線的分散程度




(a) Region 1 (b) Region 2 (c) Region 3

# Plotting the Residuals

A **residual plot** is a plot of the ($x$, residual) pairs—that is, of the pairs $(x_1, y_1 - \hat{y}_1), (x_2, y_2 - \hat{y}_2), \ldots, (x_n, y_n - \hat{y}_n)$—or of the residuals versus predicted values—the pairs $(\hat{y}_1, y_1 - \hat{y}_1), \ldots, (\hat{y}_n, y_n - \hat{y}_n)$.

理想的迴歸關係式之 Residual plot 並無明顯的殘差值之分布趨勢

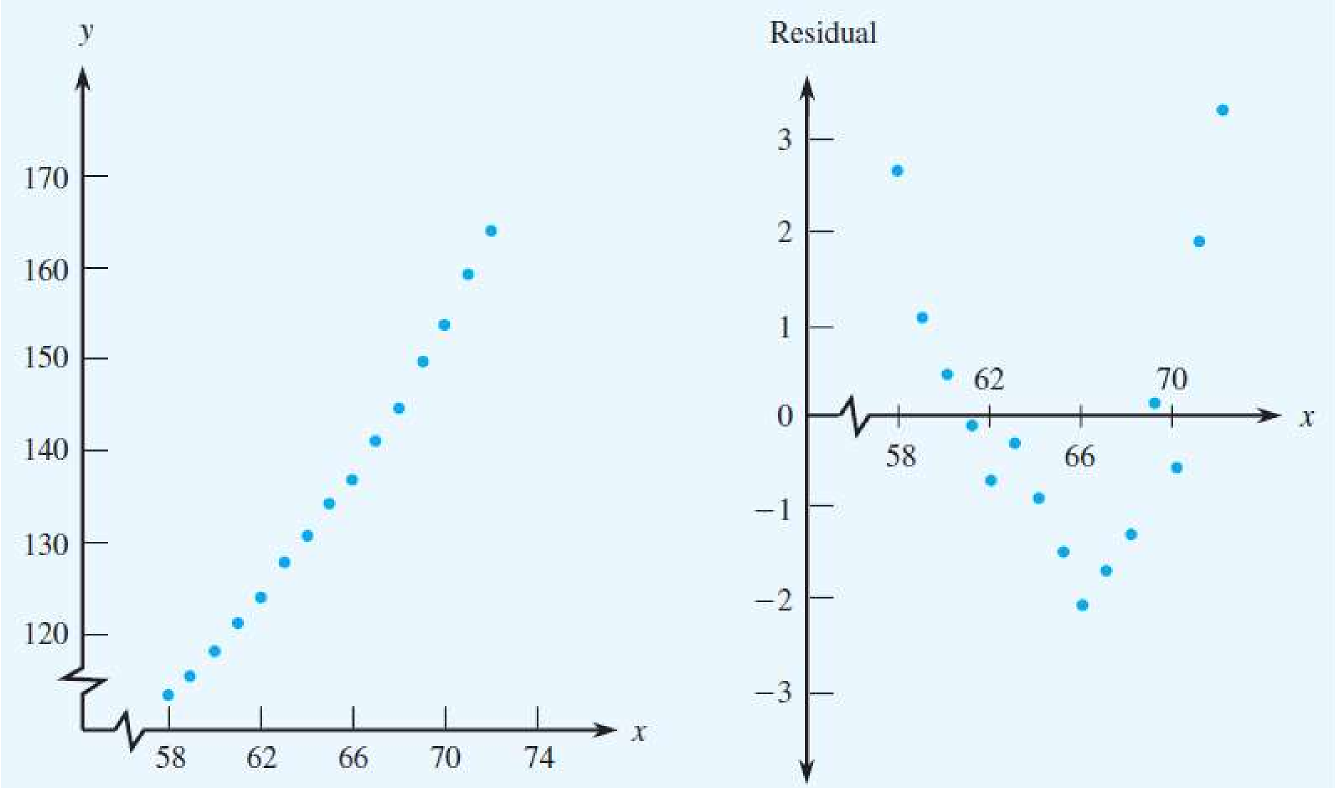
# Plotting the Residuals

應變數(Y): 美國女性體重[lb]

自變數(X): 身高 [inch]

資料個數: 15筆

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *x*: | 58 | 59 | 60 | 61 | 62 | 63 | 64 | 65 |
| *y*: | 113 | 115 | 118 | 121 | 124 | 128 | 131 | 134 |
| *x*: | 66 | 67 | 68 | 69 | 70 | 71 | 72 | |
| *y*: | 137 | 141 | 145 | 150 | 153 | 159 | 164 | |



(a) (b)

# Resistant Lines

迴歸線容易受到單一不好的觀測值影響迴歸結果。此類情況，**可以使用權重法**線性迴歸

# 補充: 檢查偏離值與權重式迴歸分析

透過殘差值定義各別資料點(x,y)的權重，並將所有資料點乘上對應的權重值

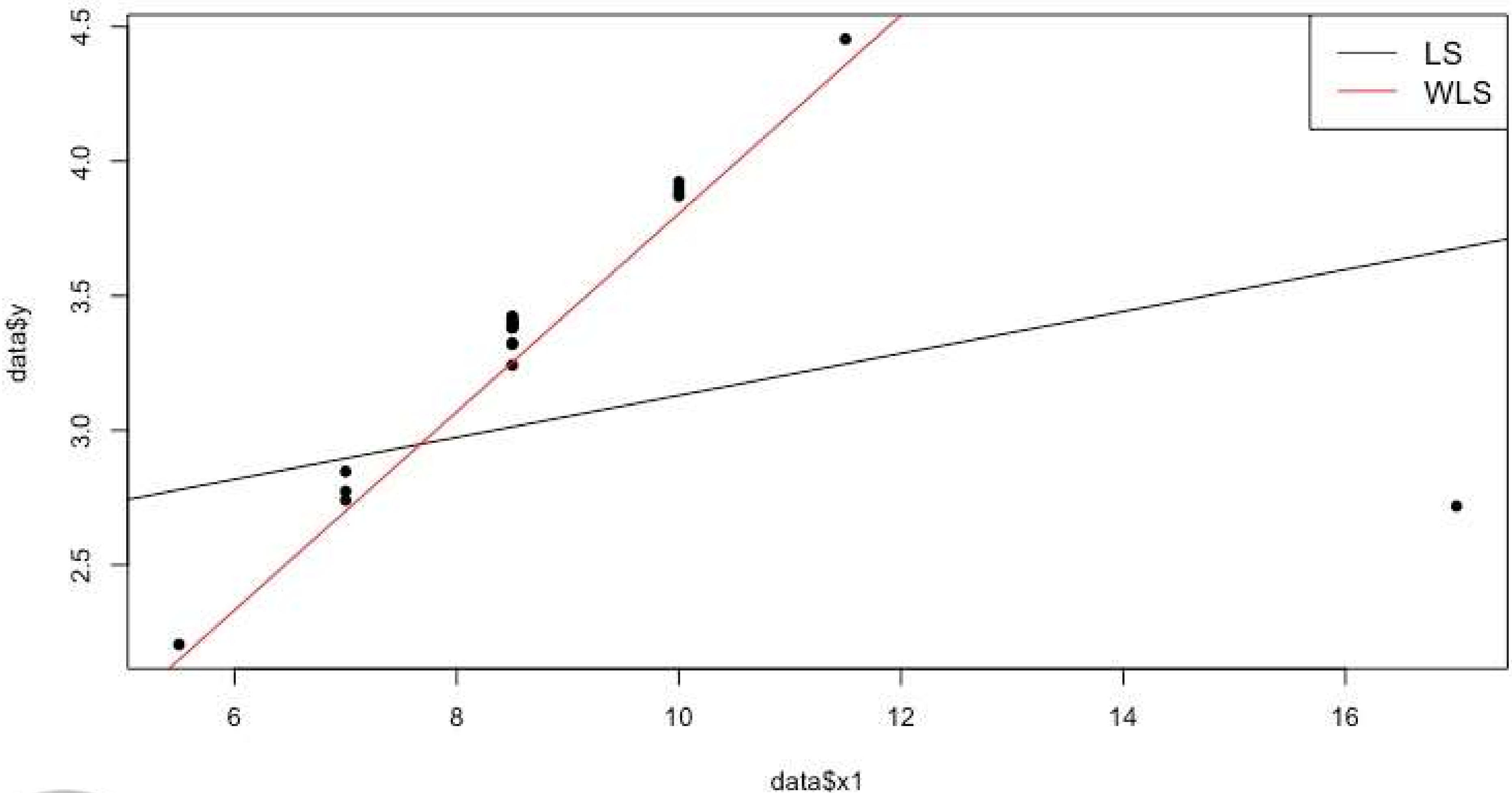
$$y_i = a + bx_i$$

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix}$$

$$\begin{bmatrix} w_1 1 & w_1 x_1 \\ w_2 1 & w_2 x_2 \\ . & . \\ . & . \\ w_n 1 & w_n x_n \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} w_1 y_1 \\ w_2 y_2 \\ . \\ . \\ w_n y_n \end{bmatrix}$$

# 補充: 檢查偏離值與權重式迴歸分析

透過乘上權重值的資料點再進行回歸分析




The R Project for Statistical Computing

# R: Fitting a Straight Line

lm(y~x).

lm(y~x+0).

anova().

# R: Fitting a Straight Line

```
Call:
lm(formula = OSA ~ Width, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.60942 -0.19875 -0.01902  0.21727  0.66378

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.3977     0.1680  -2.367   0.0251 *
Width         3.0800     0.1506  20.453   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.308 on 28 degrees of freedom
Multiple R-squared:  0.9373,	Adjusted R-squared:  0.935
F-statistic: 418.3 on 1 and 28 DF,  p-value: < 2.2e-16
```

Intercept: -0.3977
Slope: 3.0800
se: 0.308
$r^2$: 0.9373

# R: Fitting a Straight Line

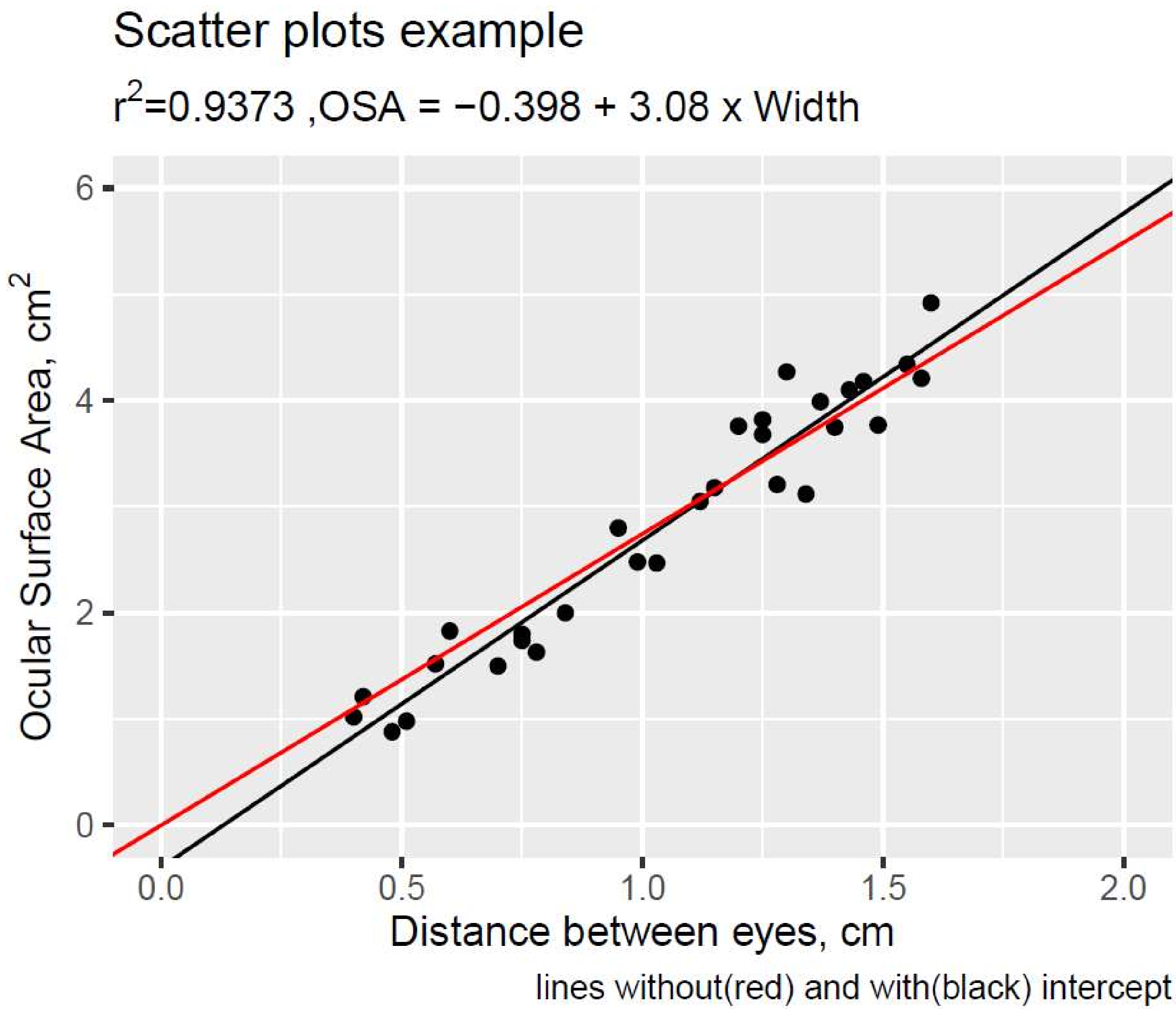
```
Analysis of Variance Table

Response: OSA
          Df Sum Sq Mean Sq F value    Pr(>F)
Width      1 39.686  39.686  418.32 < 2.2e-16 ***
Residuals 28  2.656   0.095
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SSE=2.656
SSR=39.686

$$SSTo = SSE + SSR$$

$$r^2 = 1 - \frac{SSE}{SSTo} \quad S_e = \sqrt{\frac{SSE}{n-2}}$$

# R: Fitting a Straight Line



# TRY it in R

# R: Fitting a Straight Line

R_regression_a.R

# Nonlinear Relationships

- Power Transformations

  非線性 -> 線性化...
- Fitting a Polynomial Function

  多項式曲線(非線性)
- Smoothing a Scatterplot

  平化化數據 -> 突顯資料趨勢關係

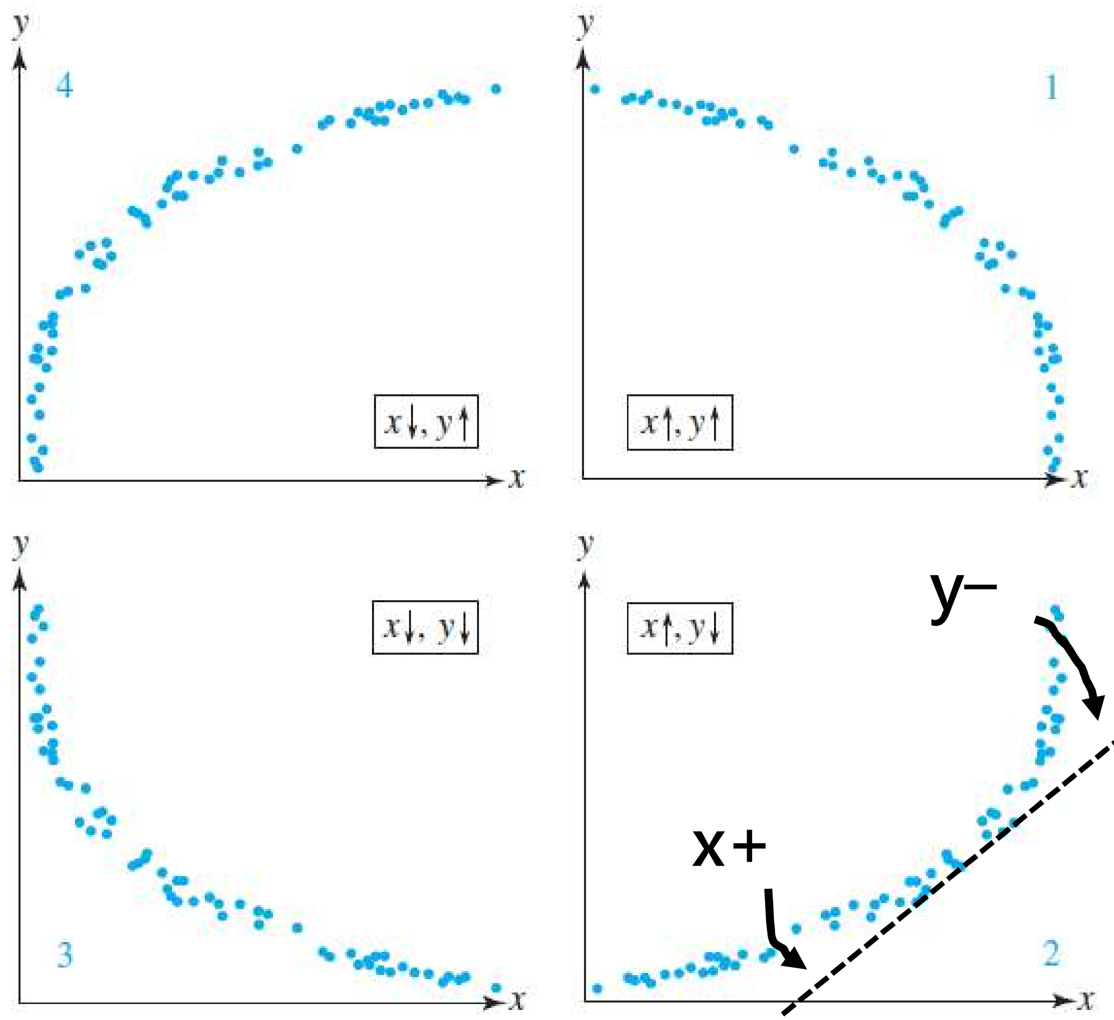
# Power Transformations

- Suppose the general pattern in a scatterplot is **curved** and **monotonic** (i.e., strictly increasing/decreasing); it is often possible to find a **power transformation** for $x$ or $y$.
- By a power transformation, we mean the use of exponents $p$ and $q$ such that the transformed values are $x' = x^p$ and/or $y' = y^q$
- The relevant scatterplot is of the $(x', y')$ pairs.

# Power Transformations

Power transformation ladder:
Transformed value = (original value)$^{\text{POWER}}$

| Power | Transformed value | Name |
|---|---|---|
| 3 | (Original value)$^3$ | Cube |
| 2 | (Original value)$^2$ | Square |
| 1 | Original value | No transformation |
| $\frac{1}{2}$ | $\sqrt{\text{Original value}}$ | Square root |
| $\frac{1}{3}$ | $\sqrt[3]{\text{Original value}}$ | Cube root |
| 0 | Log(original value) | Logarithm |
| –1 | 1/(original value) | Reciprocal |



# Power Transformations

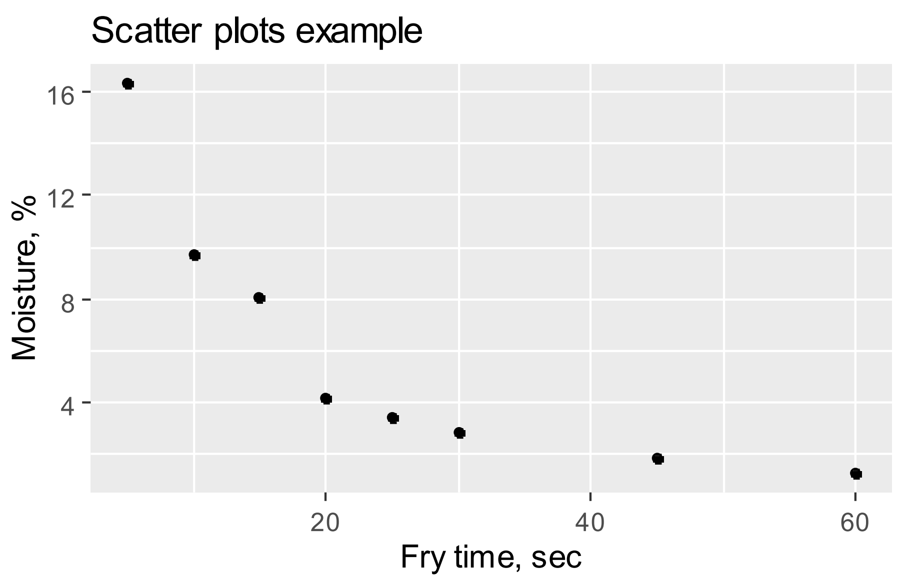
應變數(Y): 玉米脆片的溼度[%]

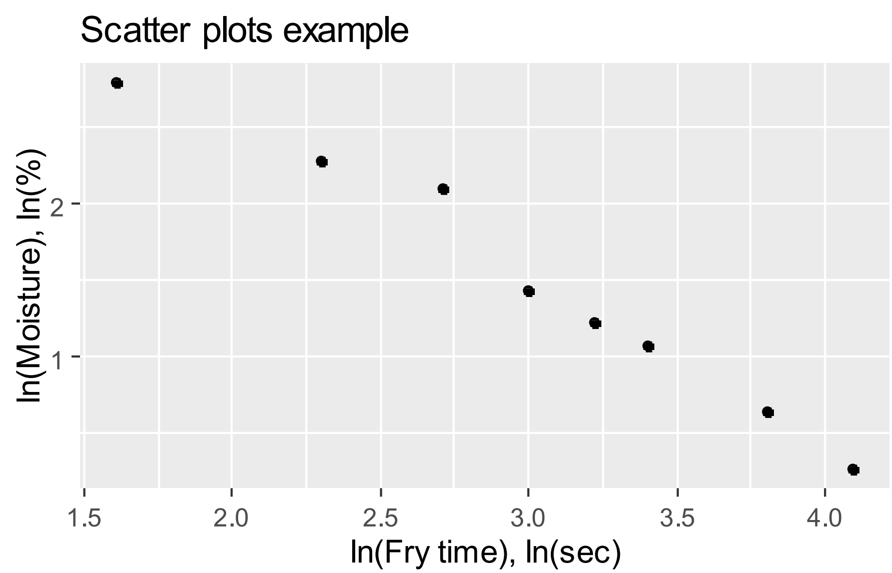
自變數(X): 油炸時間 [sec]

資料個數: 8筆, data_power.xlsx

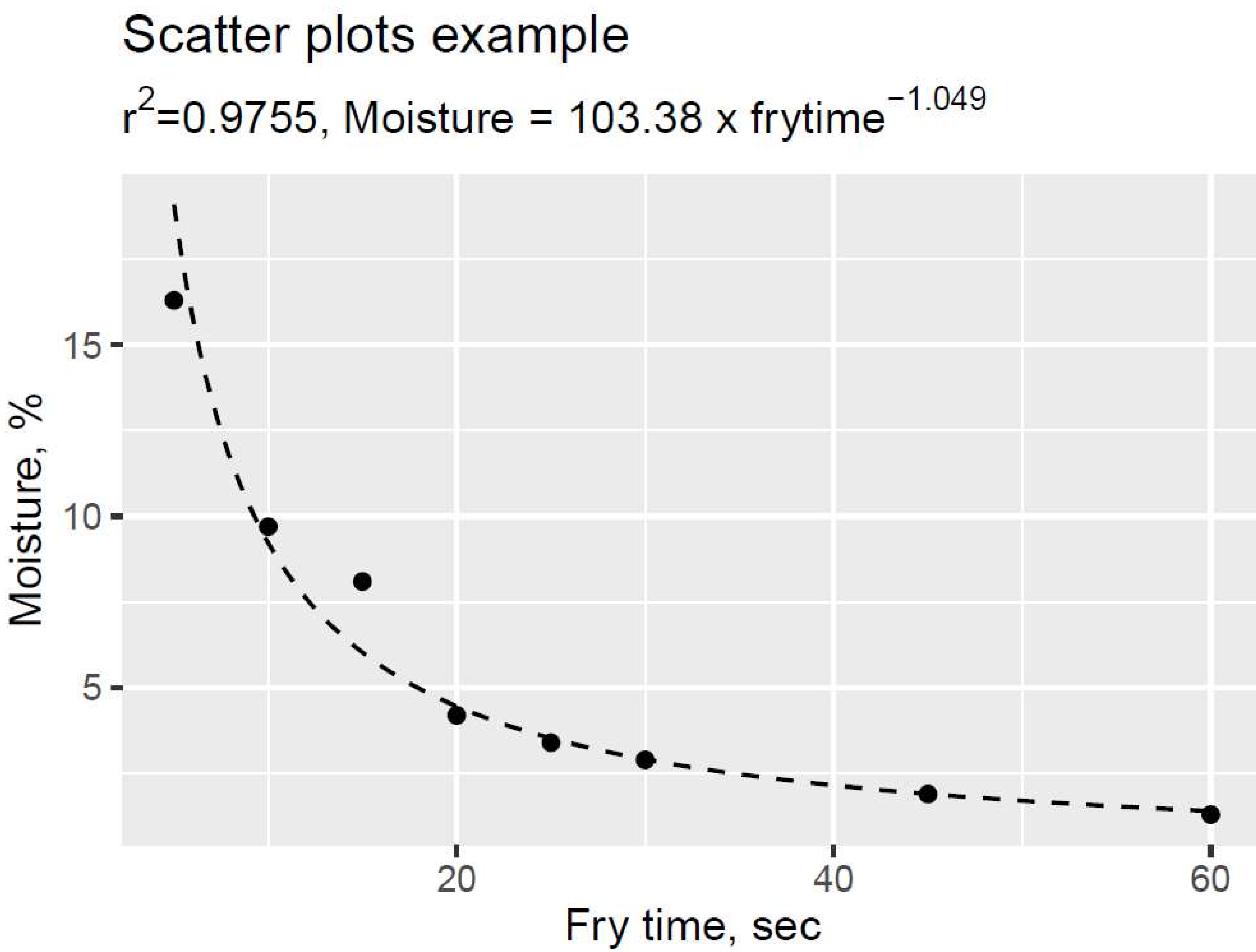
Moisture = a x time $^{b}$

ln(Moi) = ln(a)+b x ln(time)



# R: Power Transformation



TRY

it

in

R

# R: Power Transformation

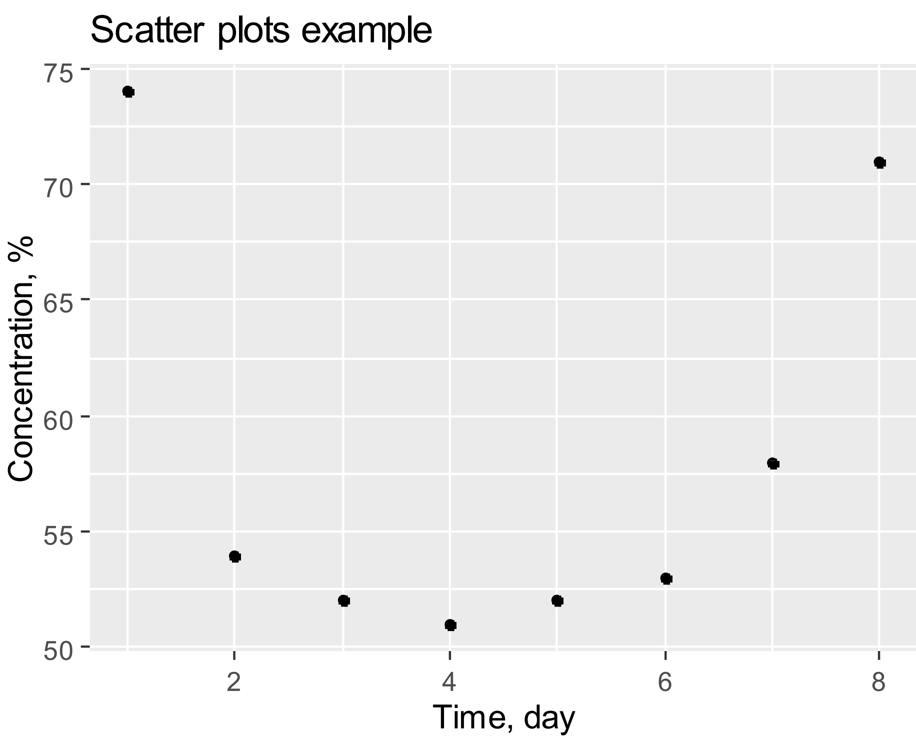
R_regression_b.R

# Fitting a Polynomial Function

應變數(Y): 葡萄糖濃度[%]

自變數(X): 發酵時間[day]

資料個數: 8筆, data_conc.xlsx




$$y = a + b_1 x + b_2 x^2$$

$$g\left(\tilde{a}, \widetilde{b_1}, \widetilde{b_2}\right) = \sum_{i=1}^{n}\left[y_i - \left(\tilde{a} + \widetilde{b_1}x + \widetilde{b_2}x^2\right)\right]^2$$

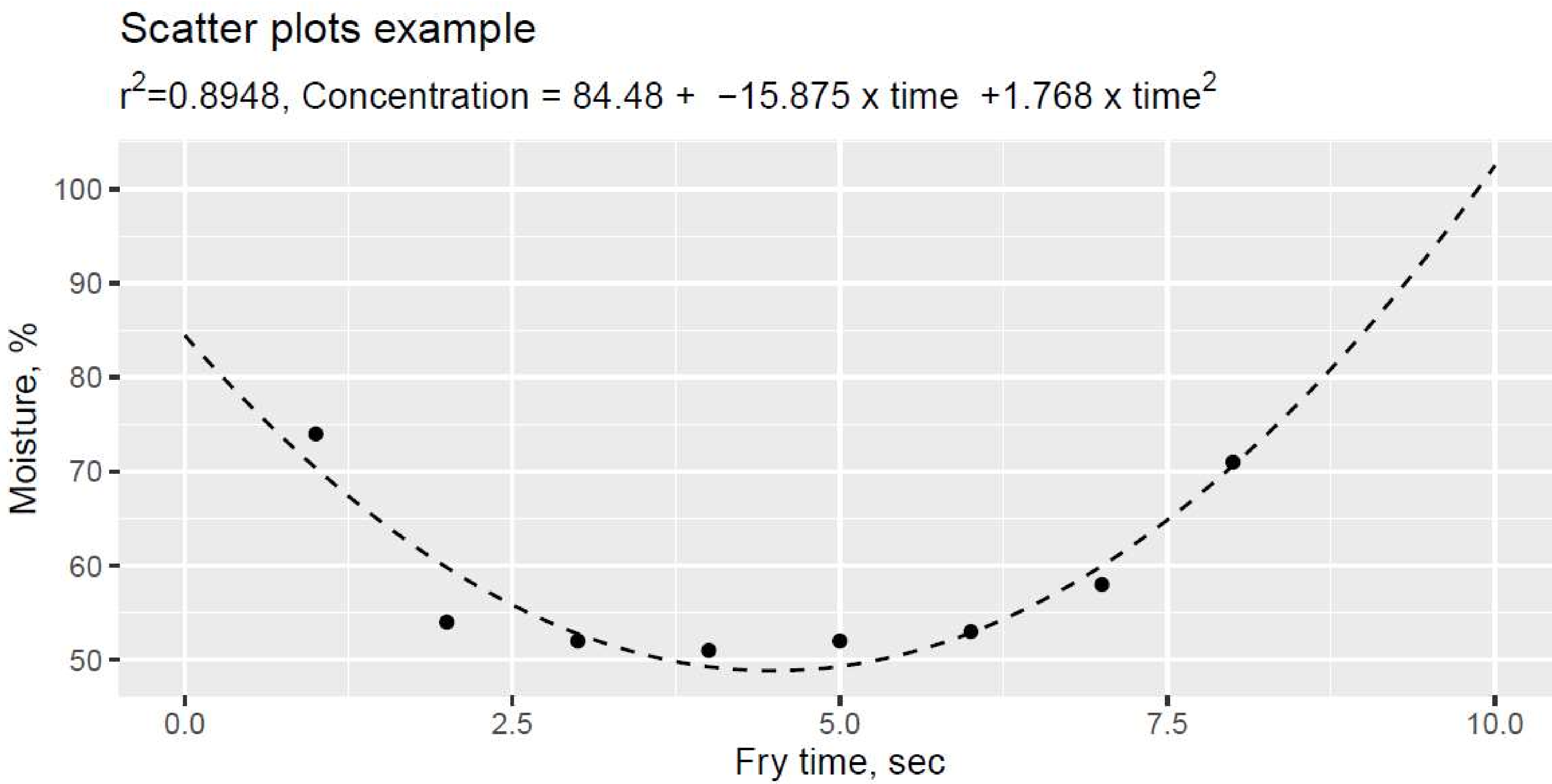
# R: Fitting a Polynomial Function

lm(y~x+I(x^2)).

# R: Fitting a Polynomial Function

```
Residuals:
      1       2       3       4       5       6 
 3.6250 -5.8036 -0.7679  1.7321  2.6964  0.1250 
      7       8 
-1.9821  0.3750 

Coefficients:
            Estimate Std. Error t value
(Intercept)  84.4821     4.9036  17.229
time        -15.8750     2.5001  -6.350
I(time^2)     1.7679     0.2712   6.519
            Pr(>|t|)    
(Intercept) 1.21e-05 ***
time         0.00143 ** 
I(time^2)    0.00127 ** 
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.515 on 5 degrees of freedom
Multiple R-squared:  0.8948,	Adjusted R-squared:  0.8527 
```

# R: Fitting a Polynomial Function




Polynomial function

# TRY
# it
# in
# R

# R: Fitting a Polynomial Function

R_regression_c.R

# Smoothing a Scatterplot

## Locally Weighted Scatterplot Smoother

**LOWESS (or LOESS) method:**

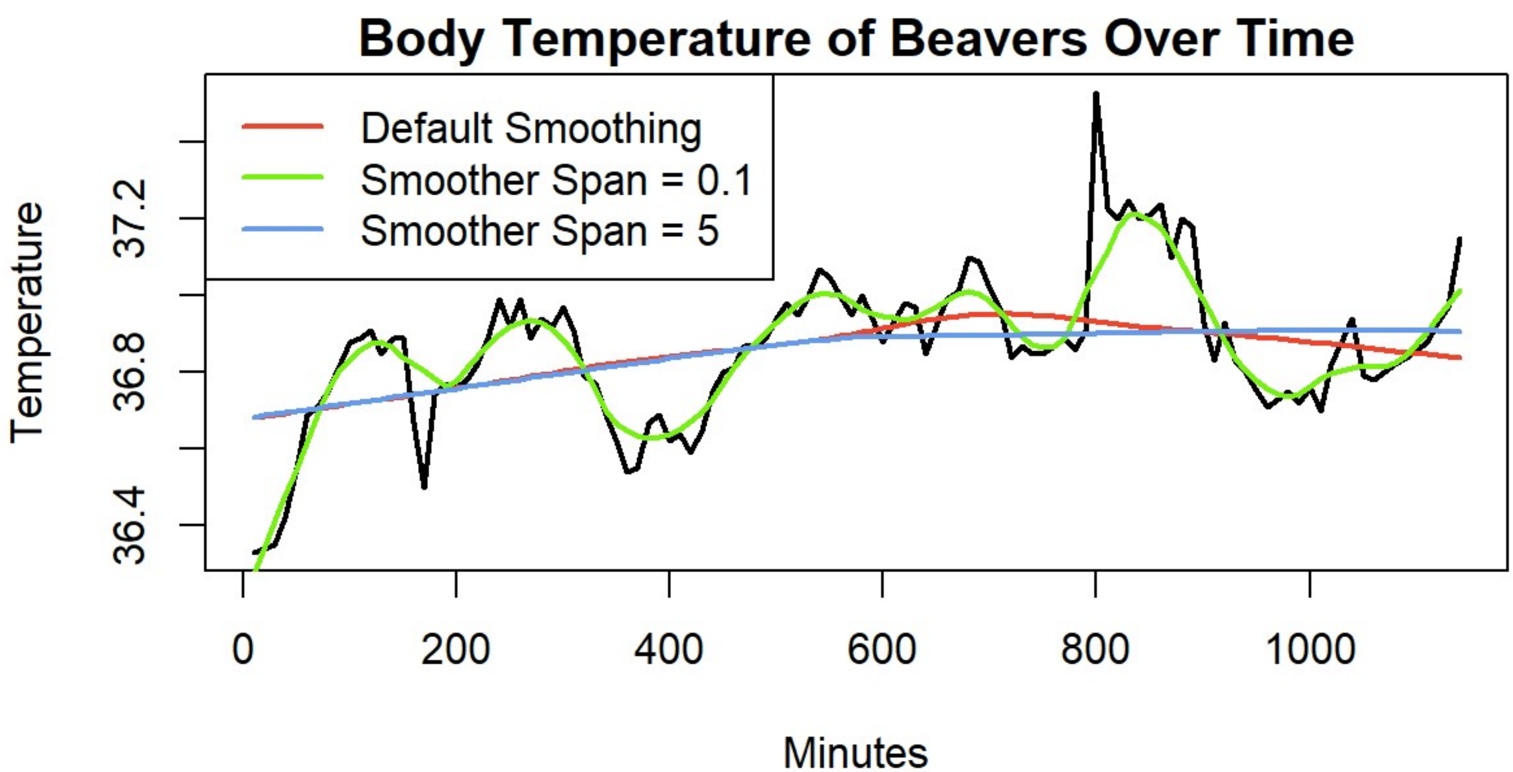
- Let $(x^*, y^*)$ denote a particular one of the $n$ $(x, y)$ pairs in the sample.
- The value corresponding to $(x^*, y^*)$ is obtained by fitting a **straight line using only a specified percentage of the data (e.g., 25%)** whose $x$ values are closest to $x^*$.
- Those with $x$ values closer to $x^*$ are more heavily weighted than those whose $x$ values are farther away.
- The height of the resulting line above $x^*$ is the fitted value .
- This process is repeated for each of the $n$ points, so $n$ different lines are fit.
- The fitted points are connected to produce a LOWESS curve.

# R語言使用Lowess Smoothing

-Use the lowess(x, y, f):

x, y: the input numeric data

f: the smooth span. Larger values give more smoothness




The R Project for Statistical Computing

# Using More than One Predictor

- Fitting a Linear Function
- Creating New Predictors from Existing Ones

# Fitting a Linear Function

- Consider fitting a relation of the form:

$$\boldsymbol{y \approx a + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k}$$

- The least squares coefficients a, $b_1$, $b_2$,…, $b_k$ are the values that minimize g:

$$\boldsymbol{g(a, b_1, \ldots, b_k) = \sum_{j=1}^{n} \left[ y_j - \left( a + b_1 x_{1j} + b_2 x_{2j} + \cdots + b_k x_{kj} \right) \right]^2}$$

# R語言使用多元線性迴歸模型

-Use the lm(y~$x_1$+$x_2$, data):

data: the input numeric data frame

$$y = a + b_1 x_1 + b_2 x_2$$

The R Project for Statistical Computing

# Creating New Predictors from Existing Ones

以迴歸分析的觀點出發，若是提供越多的Predictor應可以達到更加的擬合結果($R^2$越大)，但是，實際上應要使用最少的Predictor去達到最佳結果。

# Creating New Predictors from Existing Ones

應變數(Y): deflection
自變數($X_1$): shear span ratio
自變數($X_2$): splitting tensile strength
資料個數: 15筆

| $x_1$ | $x_2$ | $x_1x_2$ | $y$ |
|---|---|---|---|
| 2.04 | 3.55 | 7.2420 | 3.11 |
| 2.04 | 6.07 | 12.3828 | 3.26 |
| 3.06 | 3.55 | 10.8630 | 3.89 |
| 3.06 | 6.07 | 18.5742 | 10.25 |
| 4.08 | 3.55 | 14.4840 | 3.11 |
| 4.08 | 6.16 | 25.1328 | 13.48 |
| 2.06 | 3.62 | 7.4572 | 3.94 |
| 2.06 | 6.16 | 12.6896 | 3.53 |
| 3.08 | 3.62 | 11.1496 | 3.36 |
| 3.08 | 5.89 | 18.1412 | 6.49 |
| 4.11 | 3.62 | 14.8782 | 2.72 |
| 4.11 | 5.89 | 24.2079 | 12.48 |
| 2.01 | 6.18 | 12.4218 | 2.82 |
| 3.02 | 6.18 | 18.6636 | 5.19 |
| 4.03 | 6.18 | 24.9054 | 8.04 |

# Creating New Predictors from Existing Ones

Fitting $y = a + b_1x_1 + b_2x_2$

$$y = a + b_1x_1 + b_2x_2$$

$$a = -9.2744$$

$$b_1 = 2.3263$$

$$b_2 = 1.5459$$

# Creating New Predictors from Existing Ones

Including an interaction $y = a + b_1x_1 + b_2x_2 + b_3x_1x_2$

$$y = a + b_1x_1 + b_2x_2 + b_3x_1x_2$$

$$a = 17.296$$

$$b_1 = -6.373$$

$$b_2 = -3.661$$

$$b_3 = 1.708$$

# Creating New Predictors from Existing Ones

**Adding quadratic** $y = a + b_1x_1 + b_2x_2 + b_3x_1x_2 + b_4x_1^2 + b_5x_2^2$

$$y = a + b_1x_1 + b_2x_2 + b_3x_1x_2 + b_4x_1^2 + b_5x_2^2$$

$$a = -34.29104$$

$$b_1 = -6.58875$$

$$b_2 = 19.34743$$

$$b_3 = 0.06098$$

$$b_4 = -2.35896$$

$$b_5 = 1.65575$$

# Creating New Predictors from Existing Ones

Fitting $a + b_1x_1 + b_2x_2$ results in

$$\hat{y} = -9.251 + 2.322x_1 + 1.544x_2, \qquad R^2 = .576$$

Including an interaction predictor yields

$$\hat{y} = 17.279 - 6.368x_1 - 3.658x_2 + 1.707x_1x_2, \qquad R^2 = .825$$

Adding in the two quadratic predictors gives

$$\hat{y} = -34.323 - 6.568x_1 + 19.347x_2 + 1.655x_1x_2 + .058x_1^2 - 2.359x_2^2, \qquad R^2 = .845$$

# 課堂練習: 學號-姓名-ch9-Regression.R

The data give the speed of cars and distances taken to stop.

data(cars)

(1)請繪製散佈圖(scatter plot)，並計算其Pearson線性相關係數(r)，請試著說明速度(speed, x)與距離(distance, y)之間的關係

(2)請完成線性($y = a+bx$)及非線性($y = ax^b$)回歸分析，並試著說明何者較為適合描述速度(speed)與距離(distance)的關係(比較決定係數$r^2$)

[label plot: 請參考R_sampling_c1.R]

# 課堂練習: 學號-姓名-ch9-Regression.R