# Linear Regression Analysis (2): More Than One Covariate

信賴區間－用迴歸方式知道一群人的平均體重

預測區間－


杜裕康

國立台灣大學公共衛生學院
流行病學與預防醫學研究所

yukangtu@ntu.edu.tw

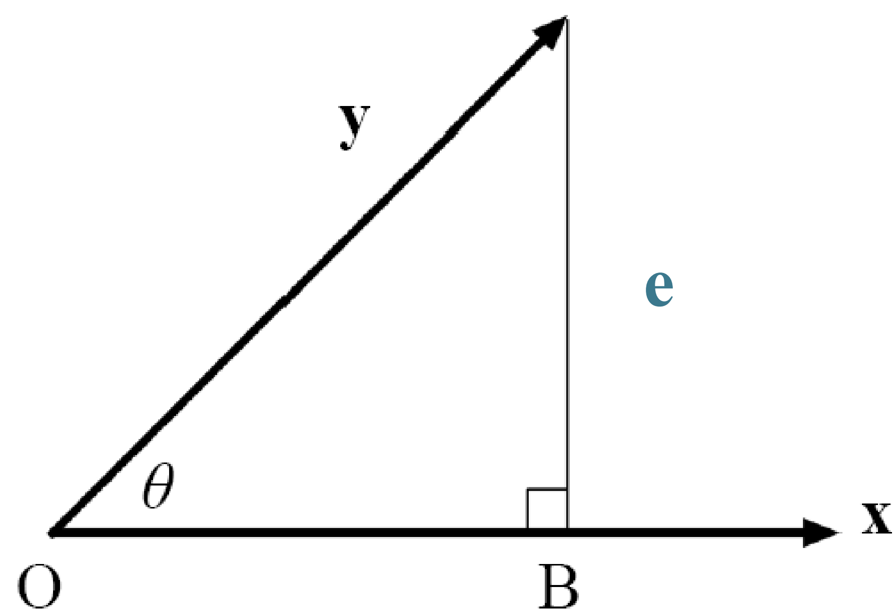
# Transformation in Vector Geometry for Linear Models

- 隨機變數 $x$ 和 $y$ 經轉換後，可以用 $n$ 維歐幾里得空間中的兩個向量 $\mathbf{x}$ 和 $\mathbf{y}$ 來表示.
- 向量的長度是其標準差 (*standard deviation*)

$$X_i^{new} = \frac{[X_i^{old} - \bar{X}]}{\sqrt{n-1}}$$

# Basics of vector geometry

- 經由剛才的轉換後，$y$ 對 $x$ 的迴歸，可以想像成 $\mathbf{y}$ 對 $\mathbf{x}$ 在向量空間中的垂直投影
- 而 $x$ 的迴歸係數，$b_x$，是 $\mathbf{y}$ 在 $\mathbf{x}$ 上投影 $\overrightarrow{OB}$ 的長度和 $\mathbf{x}$ 的長度的比例(ratio)
- $$\boldsymbol{b_x} = \frac{\left\|\overrightarrow{\boldsymbol{OB}}\right\|}{\|\mathbf{x}\|} = \frac{\|\mathbf{y}\|\boldsymbol{cos\theta}}{\|\mathbf{x}\|} = \frac{\|\mathbf{x}\|\|\mathbf{y}\|\boldsymbol{cos(\theta)}}{\|\mathbf{x}\|^2} = \frac{cov(x,y)}{Var(x)} = \frac{\mathbf{y}^T\mathbf{x}}{\|\mathbf{x}\|^2}$$




OB is the orthogonal projection of **y** on **x**

# Analysis of Variance Approach

- 當使用迴歸分析時， $Y$ 的變異亦可像在作 ANOVA 時， 分解成兩部分：

$$\Sigma\left(Y_i - \bar{Y}\right)^2 = \Sigma\left(Y_i - \hat{Y}_i\right)^2 + \Sigma\left(\hat{Y}_i - \bar{Y}\right)^2$$

$\hat{Y}_i$: $Y_i$ 的迴歸分析估計值; $\bar{Y}$: $Y_i$ 的算術平均值

- 反應變數 $Y$ 的總變異用 $\Sigma\left(Y_i - \bar{Y}\right)^2$ 來表示，稱為 total sum of squares ($SST$) ， 定義為：

$$SST = \sum\left(Y_i - \bar{Y}\right)^2$$

# Analysis of Variance Approach

- 第一部分代表觀測值和估計值的差異， 也就是 $Y$ 的變異程度裏不能被迴歸分析解釋的部分。它的平方和稱為 **error sum of squares** ($SSE$)：

$$SSE = \sum(Y_i - \hat{Y}_i)^2$$

- $SST$ 和 $SSE$ 的差別稱為 **regression sum of squares** ($SSR$) ， 就是估計值和平均值的差異的平方和：

$$SSR = SST - SSE = \sum(\hat{Y}_i - \bar{Y})^2$$

- 所以：$SST = SSE + SSR$

# Analysis of Variance Approach

- $SST$ 的分解也伴隨了自由度(degrees of freedom)的分配
- $SST$ 總共有 $n - 1$ 個自由度， 而 $SSR$ 在簡單線性迴歸裏有 1 個自由度， 所以 $SSE$ 有 $n - 2$ 個自由度。
- 所以大多數的統計軟體也會把迴歸分析的結果用接下來的 ANOVA 表格呈現：

# Analysis of Variance Approach

- The error mean square,

$$MSE = \frac{SSE}{n-2}$$

| Source of Variation | SS | df | MS | $F$ Statistic | $p$ Value |
|---|---|---|---|---|---|
| Regression | SSR | 1 | MSR = SSR/1 | $F$ = MSR/MSE | $p$ |
| Error | SSE | $n-2$ | MSE = SSE/$(n-2)$ | | |
| Total | SST | $n-1$ | | | |

$$F = t^2 \; with \; (1, n-2) \; degrees \; of \; freedom$$

# R-squared

R平方後一定比原R小

- $R^2$ (判定係數)is the percentage of variance in y explained by x:

$$R^2 = 1 - \left[\frac{\sum\left(Y_i - (a + bX_i)\right)^2}{\sum\left(Y_i - \bar{Y}\right)^2}\right] = 1 - \left[\frac{\sum\left(Y_i - \hat{Y}\right)^2}{\sum\left(Y_i - \bar{Y}\right)^2}\right] = \frac{SSR}{SST}$$

- So $R^2$ is between 0 and 1
- Adjusted $R^2$ takes into account the number of explanatory variables in the model:

$$R^2_{adj} = 1 - \left[\frac{\sum\left(Y_i - (a + bX_i)\right)^2/(n-k-1)}{\sum\left(Y_i - \bar{Y}\right)^2/(n-1)}\right]$$

# Estimation of Parameters 算截距下的標準誤

- The standard errors for the estimated *a* ($\hat{a}$):

$$se(\hat{a}) = s_{y|x}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}}$$

- The se for the estimated *b* ($\hat{b}$):

若上面數值不變，下方平方和（數值）越大，斜率就越精準

$$se(\hat{b}) = \frac{S_{y|x}}{\sqrt{\sum(X_i - \bar{X})^2}}$$

$$S_{y|x} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum(Y_i - (a + bX_i))^2}{n-2}}$$

# Estimation of Parameters

- 對迴歸係數 $b$ 作統計檢定：

$$t = \frac{\hat{b}}{se_{\hat{b}}}$$

**用T檢定**

服從 $t$-distribution with $n$ – 2 degrees of freedom

虛無假設：$b = 0$

**雙尾檢定**

對立假設：$b \neq 0$

- 95% confidence interval:

95% CI = $b$ – 1.96*se to $b$ + 1.96*se, when $n \to \infty$

# $Y$ 的信賴區間

- 當 $X$ 為 $x$， $Y$ 的期望值 $\hat{y} = \hat{a} + \hat{b}x$
- 而我們可以計算Y 的期望值 $\hat{y}$ 的信賴區間
- $\hat{y}$ 的標準誤(*standard error*)為:

**X，不是固定的，要帶入不同的身高**

$$se(\hat{y}) = S_{y|x}\sqrt{\left[\frac{1}{n} + \frac{(x - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]}$$

- 因為信賴區間受 $x$ 的影響，所以不是固定的

**距離X值的信賴區間越大**

**中間大，兩邊呈喇叭狀**

# $Y$ 的預測區間　看資料的分佈

- 假設我們從母群體中抽出一個 $X$ 為 $x$ 的個体體，它的 $Y$ 值的預測值為 $\tilde{y}$：

$$\tilde{y} = \hat{a} + \hat{b}x$$

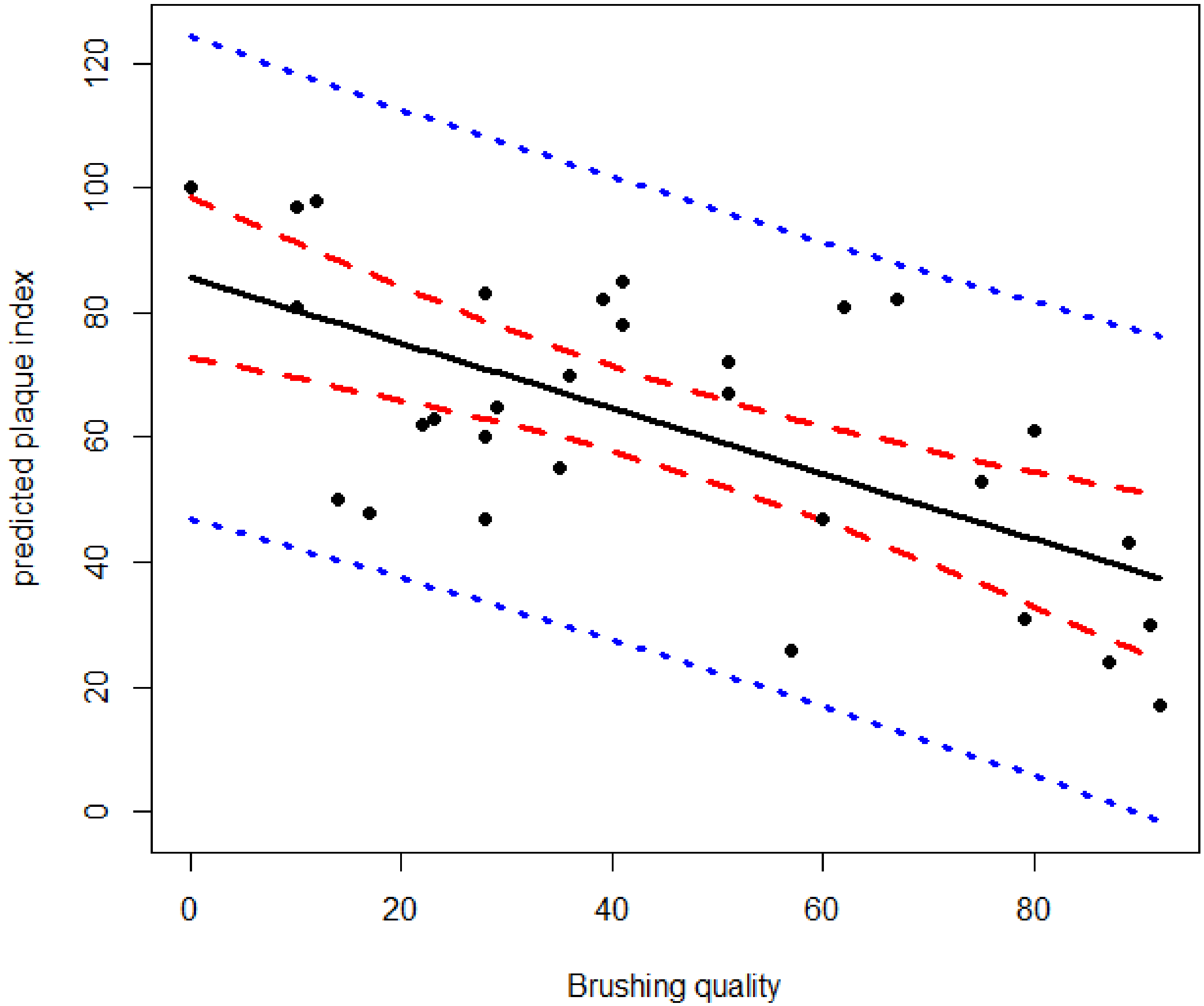
- $\tilde{y}$ 的可能範圍，稱為 $Y$ 的預測區間 (prediction interval)

多了1,故會比信賴區間大

- 而 $\tilde{y}$ 的標準誤(*standard error*)為:

$$se(\tilde{y}) = S_{y|x}\sqrt{\left[1 + \frac{1}{n} + \frac{(x-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right]}$$

# Confidence Interval and Prediction Interval Band




預測區間

信賴區間

# Presentation

- 呈現迴歸分析的結果時，應報告截距項和迴歸係數的值和它們的信賴區間
- 報告它們的 *p* 值（記得截距項和迴歸係數的虛無假設為何？ ）
- 通常我們在意的是迴歸係數，也就是斜率
- $R^2$（判定係數）是指依變項 *Y* 的變異程度能被解釋變項 *X* 解釋的百分比。

**R平方越大，越能解釋變項X的百分比**

# Multiple Regression (複迴歸)

- 簡單迴歸和複迴歸之間的最大差別，是在複迴歸裏，有超過一個以上的解釋變數，

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_pX_p + e$$

- 而我們希望用較多的解釋變數來說明或預測 $Y$，以增加模型的解釋能力，例如 $R^2$

# OLS Multiple Regression

- For $Y = a + \sum b_j X_j + e$, OLS regression tries to minimize the error sum of squares:

$$\sum \left(Y_i - \left(a + \sum b_j X_{ij}\right)\right)^2$$

- To obtain $b_j$, we need a little bit of linear algebra

$$\mathbf{y} = \mathbf{X}\mathbf{b}$$
$$\mathbf{X}^{\mathbf{T}}\mathbf{y} = \mathbf{X}^{\mathbf{T}}\mathbf{X}\mathbf{b}$$
$$\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} = \mathbf{b}$$

用舉証的方式證明

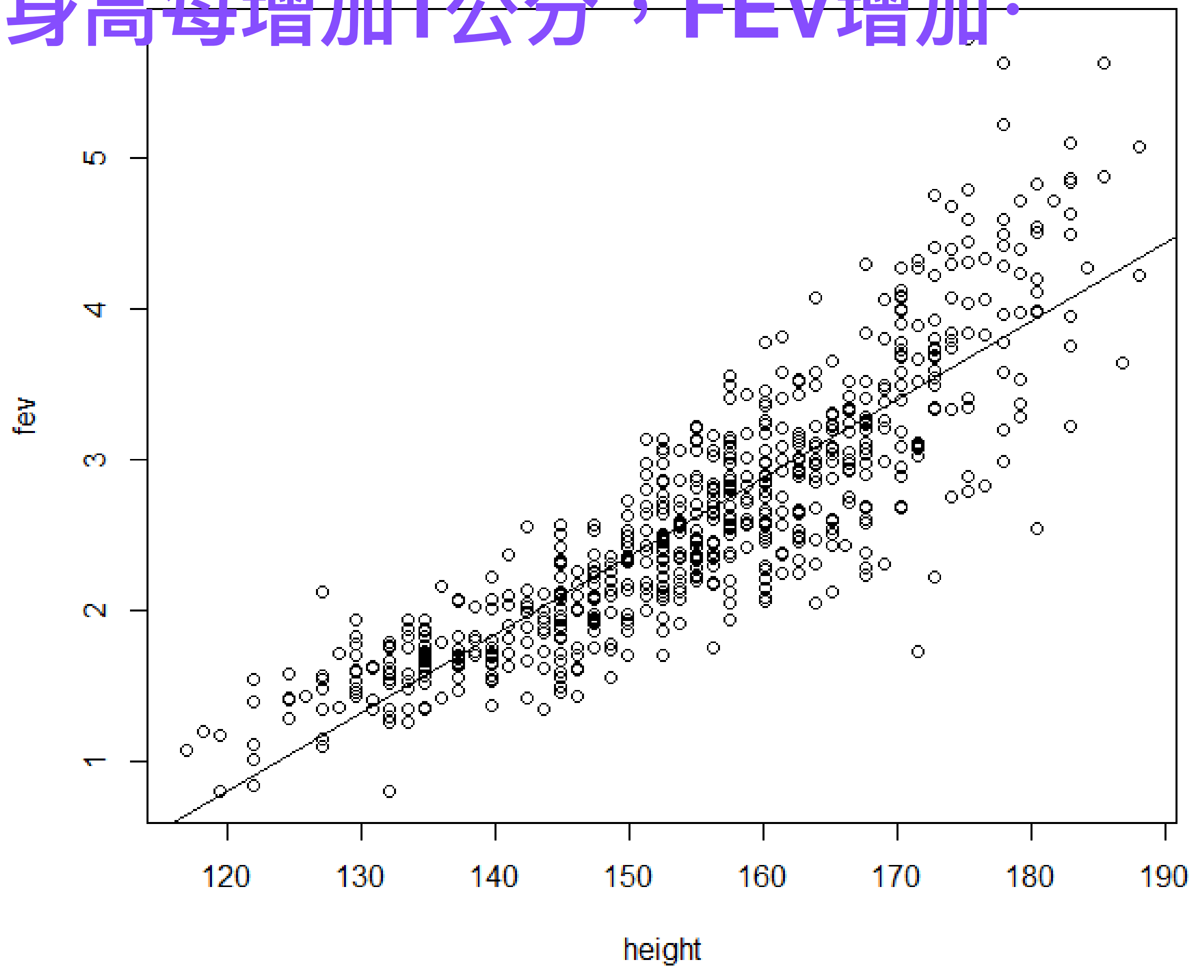
# Example: 654個 3 到 18 歲的兒童和青少年身高 (height)和最大呼氣量的關係：

| fev | height |
|---|---|
| 1.708 | 144.78 |
| 1.724 | 171.45 |
| 1.72 | 138.43 |
| 1.558 | 134.62 |
| 1.895 | 144.78 |
| 2.336 | 154.94 |
| 1.919 | 147.32 |
| 1.415 | 142.24 |
| 1.987 | 148.59 |
| 1.942 | 152.4 |
| … | … |

**FEV: Forced Expiratory Volume**

$$fev = -5.43 + 0.05height + e$$

身高每增加1公分，FEV增加·



- 現在我們想知道年齡(age) 、 身高(height)和最大呼氣量(fev)的關係：

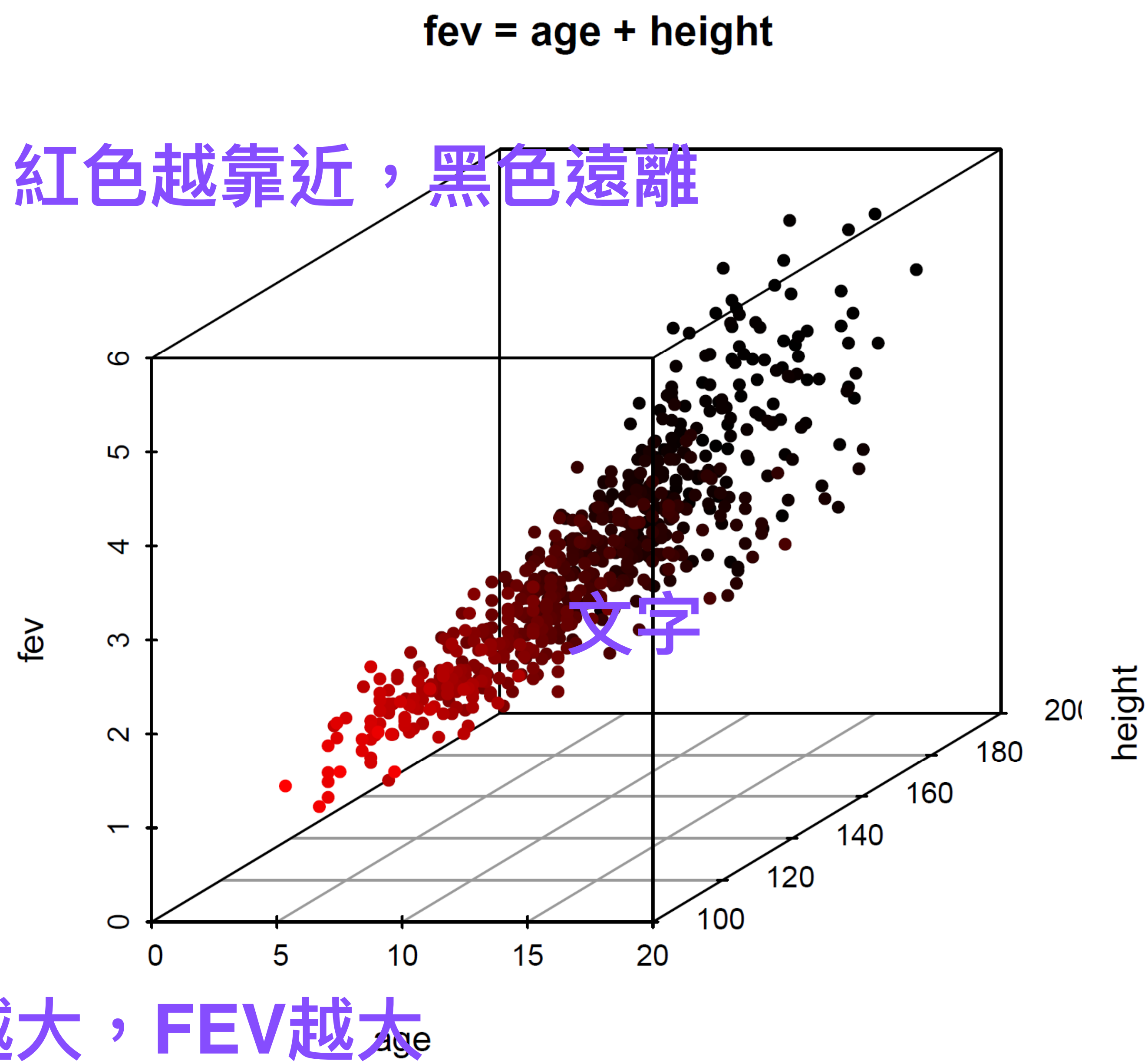
這是平面的公式

$$fev = b_0 + b_1 height + b_2 age + e$$

| fev | height | age |
|---|---|---|
| 1.708 | 144.78 | 9 |
| 1.724 | 171.45 | 8 |
| 1.72 | 138.43 | 7 |
| 1.558 | 134.62 | 9 |
| 1.895 | 144.78 | 9 |
| 2.336 | 154.94 | 8 |
| 1.919 | 147.32 | 6 |
| 1.415 | 142.24 | 6 |
| 1.987 | 148.59 | 8 |
| 1.942 | 152.4 | 9 |
| 1.602 | 134.62 | 6 |
| 1.735 | 137.16 | 8 |
| 2.193 | 148.59 | 8 |
| 2.118 | 153.67 | 8 |
| 2.258 | 147.32 | 8 |
| …… | | |

年龄越大，FEV越大

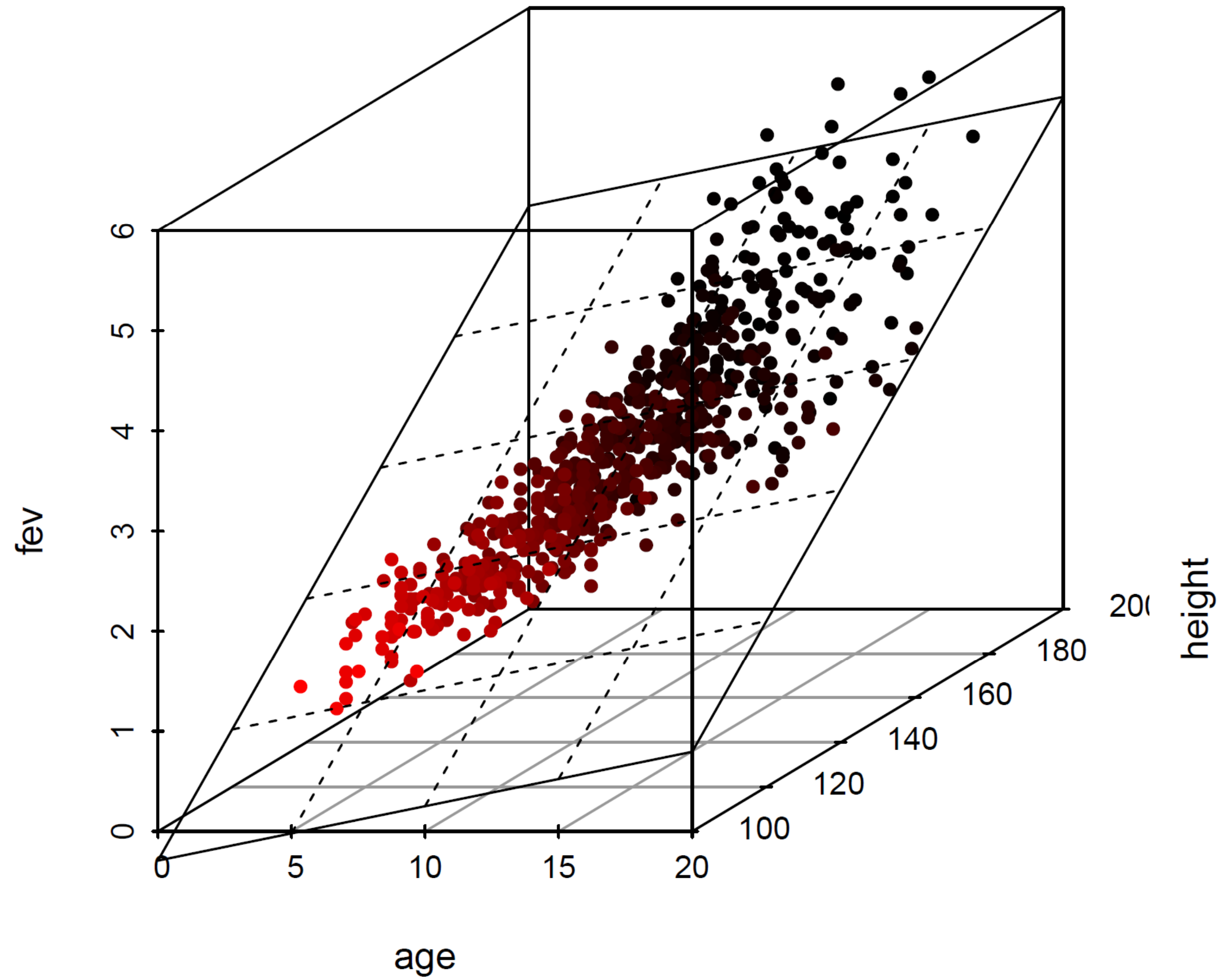
紅色越靠近，黑色遠離




文字

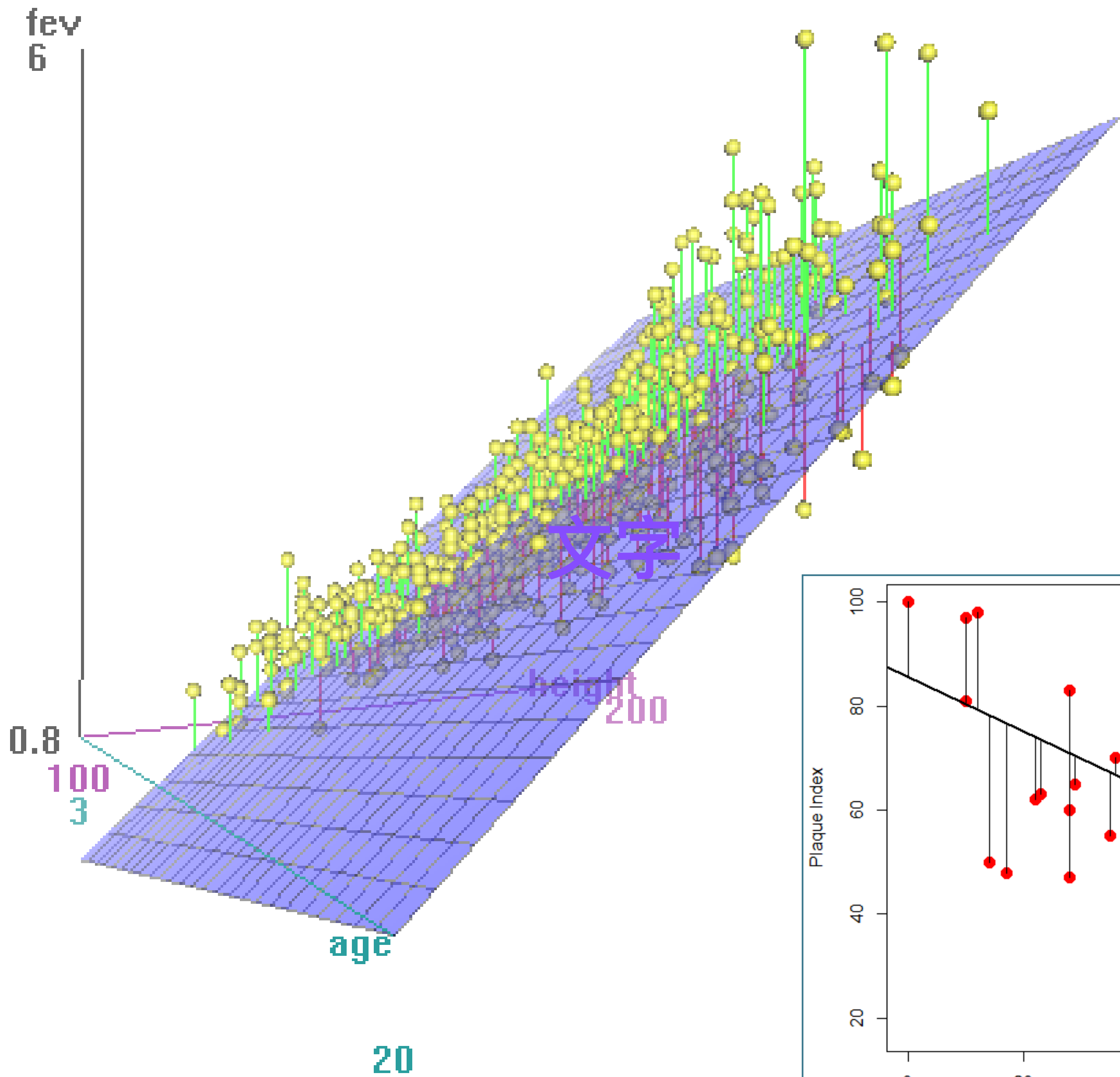
年齡越大，FEV越大

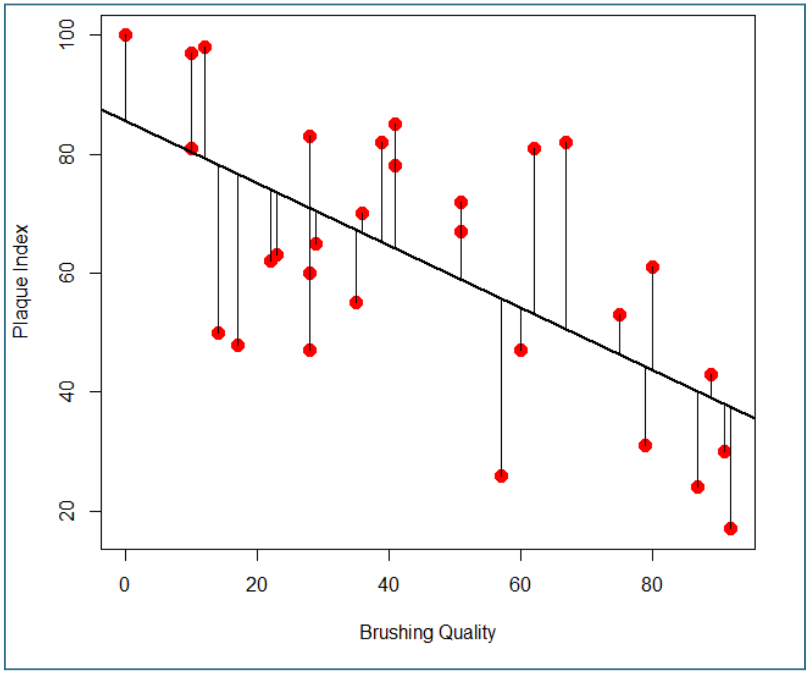
## fev = age + height



點到平面的距離之總和

在四維的空間裡找出最好



# R command

當一個人年齡 / 身高為0時其FEV為－4.61

```
> lm1<-lm(fev~height+age,data=fev)
> summary(lm1)
```

如果身高都相同，我比你多一歲身高多0.043公分，其FEV會增加

```
Call:
lm(formula = fev ~ height + age, data = fev)

Residuals:
     Min       1Q   Median       3Q      Max
-1.50533 -0.25657 -0.01184  0.24575  2.01914

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.610466   0.224271 -20.558  < 2e-16 ***
height       0.043194   0.001857  23.263  < 2e-16 ***
age          0.054281   0.009106   5.961 4.11e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4197 on 651 degrees of freedom
Multiple R-squared:  0.7664,    Adjusted R-squared:  0.7657
F-statistic:  1068 on 2 and 651 DF,  p-value: < 2.2e-16
```

除了看個別變項，還要看整體模型

# Analysis of Variance Approach

- 還記得 *SST* (total sum of squares) 定義為：

$$SST = \sum\left(Y_i - \bar{Y}\right)^2$$

- 而 *SSR* (regression sum of squares)， 也就是 *Y* 的變異裏**可以**被迴歸分析解釋的部分：

$$SSR = \sum\left(\hat{Y}_i - \bar{Y}\right)^2$$

- 而*SSE* (error sum of squares)， 是 Y 的變異裏**不能**被迴歸分析解釋的部分，也就是觀測值和估計值之間的差別：

$$SSE = \sum\left(Y_i - \hat{Y}_i\right)^2$$

# Analysis of Variance Approach

- 之前提到過：$\boldsymbol{SST = SSE + SSR}$
- $SST$ 的分解也伴隨了自由度(degrees of freedom)的分配
- $SST$ 總共有 $n-1$ 個自由度， 而 $SSR$ 在複迴歸裏有 $k$ 個自由度（$k$ 是解釋變數的數目）， 所以 $SSE$ 有 $n-k-1$ 個自由度。
- 統計軟體也會把複迴歸分析的結果用接下來的 ANOVA 表格呈現：

# Analysis of Variance Approach

| Source of Variation | SS | df | MS | $F$ Statistic | $p$ Value |
|---|---|---|---|---|---|
| Regression | SSR | $k$ | $\text{MSR} = \text{SSR}/k$ | $F = \text{MSR}/\text{MSE}$ | $p$ |
| Error | SSE | $n-k-1$ | $\text{MSE} = \text{SSE}/(n-k-1)$ | | |
| Total | SST | $n-1$ | | | |

- $R^2$ (判定係數) is the percentage of variance in y explained by x:

$$R^2 = \frac{SSR}{SST}$$

- So $R^2$ is between 0 and 1

# Analysis of Variance Approach

迴歸一樣可以用ANOVA來解釋

```
Analysis of Variance Table

Response: fev
           Df Sum Sq Mean Sq  F value    Pr(>F)    
height      1 369.99  369.99 2100.380 < 2.2e-16 ***
age         1   6.26    6.26   35.532 4.112e-09 ***
Residuals 651 114.67    0.18
```

在調整年齡（其他變數）之後，每增加1歲FEV增加

$$R^2 = \frac{369.99 + 6.26}{369.99 + 6.26 + 114.67} = 0.7664 = 76.64\%$$

# Interpretation

$$fev = -4.61 + 0.043height + 0.054age + e$$

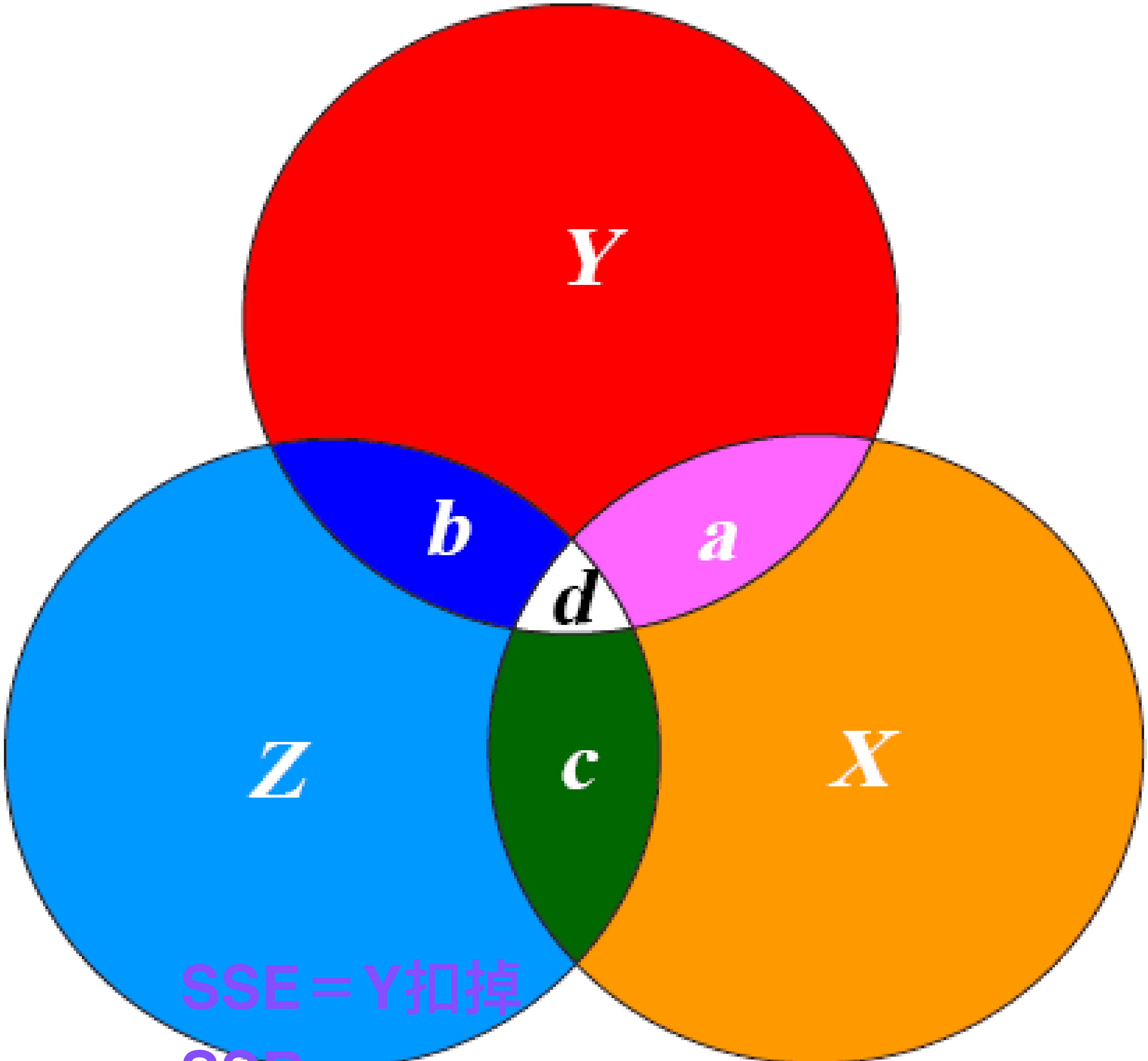
$$\widehat{fev} = -4.61 + 0.043height + 0.054age$$

- 當調整年齡之後，身高每增加 1 公分，最大呼氣量增加 0.043 公升
- 當調整身高之後，年齡每增加 1 歲 ，最大呼氣量增加 0.054 公升
- 截距項 -4.61 要怎麼解釋？

解釋
結果

# Venn Diagram

**3個圈＝3個變數，圈的大小代表，重疊越大，其關係越緊密**

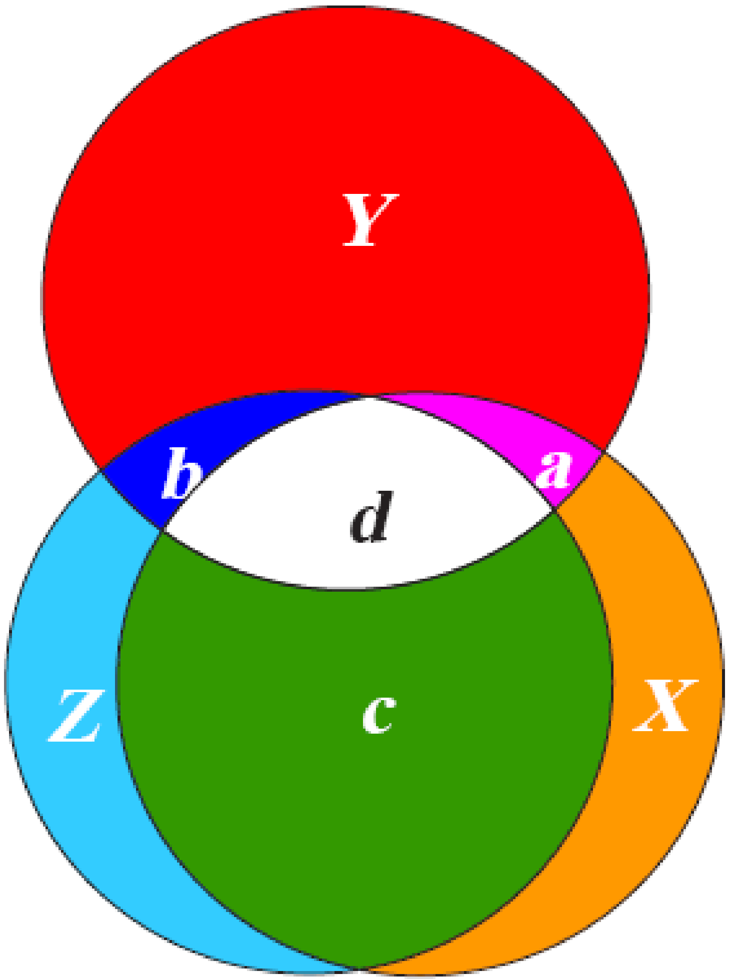



SSE＝Y扣掉

SSR＝

- 當 Y 對 X 和 Z 做迴歸分析時， X 的迴歸係數取決於 X 對 Y 的獨立解釋，也就是考慮了其它變數對 Y 的影響之後，X 對 Y的獨立影響。
- 以左圖為例：a 和 d 是Y可以被 X 解釋的部分，但是只有 a 是 X 對 Y 的獨立影響

# Venn Diagram




- 所以當 X 和 Z 對 Y 的解釋重複的部份很大的時候，X 對 Y 的獨立解釋的部份，有可能變得很小。也是加入Z之後對X的迴歸係數造成很大的改變。

**單看XYZ其貢獻不大，但一起看就很大**

# Standardized Regression Coefficient (標準化迴歸係數)

- 當所有的變數都被中心化(centered)和縮放(scaled)成變異數成為 1 的時候， 所得到的迴歸係數。
- 有些人喜歡用標準化迴歸係數來詮釋依變項和解釋變項之間關係的強度
- 當只有一個解釋變項時，標準化迴歸係數就等於相關係數， 所以它的範圍在 -1 和 1 之間。

# Standardized Regression Coefficient (標準化迴歸係數)

- 當有超過一個以上的解釋變項時，標準化迴歸係數就**不會**等於相關係數，所以它的範圍**不會**在 -1 和 1 之間。
- 要得到標準化迴歸係數(通常用希臘字母beta)，可在做迴歸分析時利用 *R* 的 scale 函數

**會把系數標準化**

```
> lm2<-lm(scale(fev)~scale(height)+scale(age),data=fev)
> summary(lm2)

Call:
lm(formula = scale(fev) ~ scale(height) + scale(age), data =
fev)

Residuals:
     Min       1Q   Median       3Q      Max
-1.73613 -0.29591 -0.01366  0.28343  2.32872

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.986e-16  1.893e-02   0.000        1
scale(height)  7.217e-01  3.102e-02  23.263  < 2e-16 ***
scale(age)     1.849e-01  3.102e-02   5.961 4.11e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4841 on 651 degrees of freedom
Multiple R-squared:  0.7664,    Adjusted R-squared:  0.7657
F-statistic:  1068 on 2 and 651 DF,  p-value: < 2.2e-16
```

## Standardized Regression Coefficient (標準化迴歸係數)

- 身高和年齡的標準化迴歸係數分別是 0.722 和 0.185
- 截距項為 0
- 標準化不影響 significance testing, model $R^2$, etc.

# Categorical Explanatory Variables (類別解釋變項)

- 像是**性別**和**種族**之類的變項屬於類別解釋變項
- 在進行迴歸分析時， 我們利用虛擬變項(dummy variables)來代表類別解釋變項
- 例如：

| Subjects | Women | Men |
|---|---|---|
| 1 | 1 | 0 |
| 2 | 0 | 1 |
| 3 | 0 | 1 |
| 4 | 1 | 0 |

# Categorical Explanatory Variables (類別解釋變項)

有K個類別，我們需要K－1個虛擬變項

- For $k$ categories, we need $k-1$ dummies

| fev | height | age | gender |
|---|---|---|---|
| 1.708 | 144.78 | 9 | 0 |
| 1.724 | 171.45 | 8 | 0 |
| 1.72 | 138.43 | 7 | 0 |
| 1.558 | 134.62 | 9 | 1 |
| 1.895 | 144.78 | 9 | 1 |
| 2.336 | 154.94 | 8 | 0 |
| 1.919 | 147.32 | 6 | 0 |
| 1.415 | 142.24 | 6 | 0 |
| 1.987 | 148.59 | 8 | 0 |
| 1.942 | 152.4 | 9 | 0 |
| 1.602 | 134.62 | 6 | 0 |
| 1.735 | 137.16 | 8 | 1 |
| 2.193 | 148.59 | 8 | 0 |
| 2.118 | 153.67 | 8 | 1 |
| 2.258 | 147.32 | 8 | 1 |
| …… | | | |

# 類別解釋變項

- 假設現在我們想知道性別(gender)和最大呼氣量(fev)的關係，其中男生設為 1 ，女生設為 0
- 模型可以寫成： $fev = b_0 + b_1 gender + e$

所有女性的FEV為2.45
GENder，男女差0.36
P值夠小－》有意義

```
> lm1<-lm(fev~gender, data=fev)
> summary(lm1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.45117    0.04759  51.505  < 2e-16 ***
gender       0.36128    0.06640   5.441  7.5e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8487 on 652 degrees of freedom
Multiple R-squared:  0.04344,   Adjusted R-squared:  0.04197
F-statistic: 29.61 on 1 and 652 DF,  p-value: 7.496e-08
```

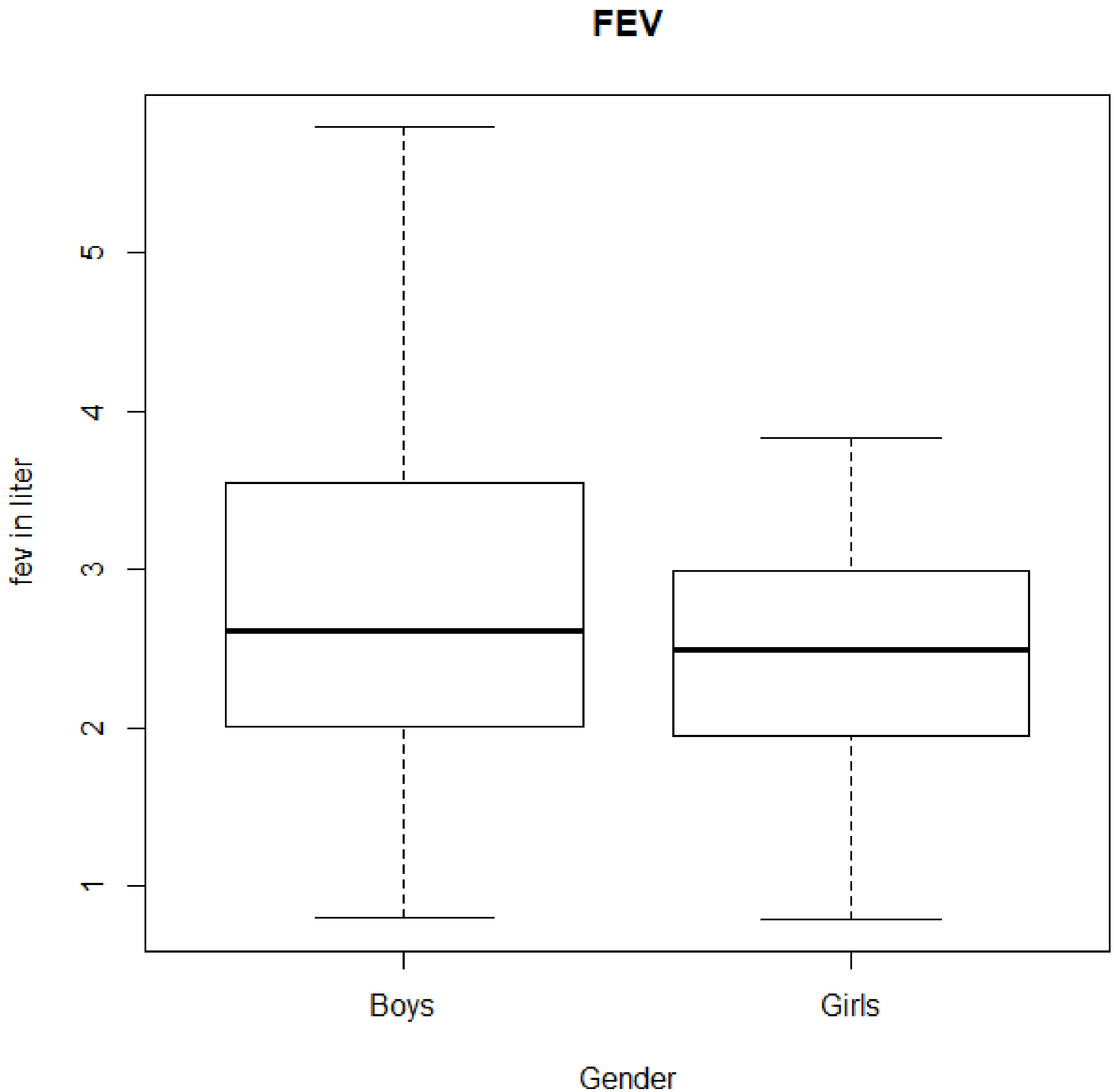
# 類別解釋變項

- 截距項 $b_0 = 2.45$ 代表女生最大呼氣量的平均值
- *gender* 的迴歸係數 $b_0 = 0.36$ 代表所有男生最大呼氣量的平均值和女生最大呼氣量的平均值**的差異**，這個差異達到統計上顯著，因為 $p = 0.000000075$

$$t = \frac{0.361}{0.066} = 5.441$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.45117    0.04759  51.505  < 2e-16 ***
gender       0.36128    0.06640   5.441  7.5e-08 ***
```

# Boxplot



# t-test 和迴歸分析的關係

- 我們也可以用兩獨立樣本的 t 檢定來比較男生和女生最大呼氣量平均值的差異，得到相同的結果

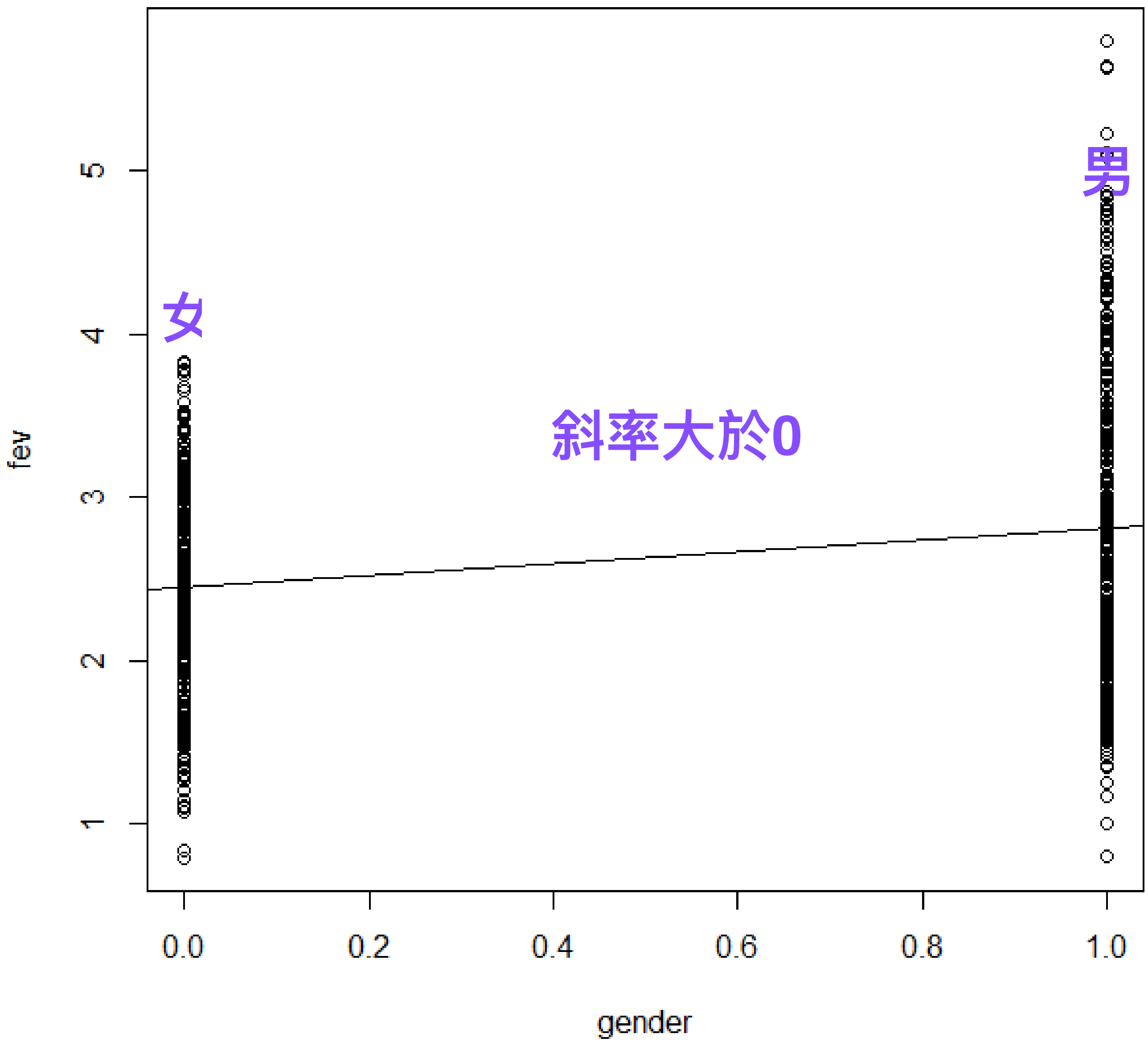
```
> t.test(fev~gender, var.equal=T, data=fev)

        Two Sample t-test

data:   fev by gender
t = -5.4412, df = 652, p-value = 7.496e-08
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -0.4916530 -0.2309002
sample estimates:
mean in group 0 mean in group 1
       2.451170        2.812446
```

# t-test 和迴歸分析的關係

```
plot(fev~gender,data=fev, main="FEV")
abline(lm1)
```



# ANOVA和迴歸分析的關係

- 假設現在我們把年齡分成三組：
  - <6 “Preschool”
  - 7-12 “Primary”
  - >12 “Secondary”
- 然後看年齡組(agegp)和最大呼氣量(fev)的關係，

# ANOVA和迴歸分析的關係

| | fev | height | age | gender | smoking | sex | agegp |
|---|---|---|---|---|---|---|---|
| 1 | 1.708 | 144.780 | 9 | 0 | 0 | Girls | Primary |
| 2 | 1.724 | 171.450 | 8 | 0 | 0 | Girls | Primary |
| 3 | 1.720 | 138.430 | 7 | 0 | 0 | Girls | Primary |
| 4 | 1.558 | 134.620 | 9 | 1 | 0 | Boys | Primary |
| 5 | 1.895 | 144.780 | 9 | 1 | 0 | Boys | Primary |
| 6 | 2.336 | 154.940 | 8 | 0 | 0 | Girls | Primary |
| 7 | 1.919 | 147.320 | 6 | 0 | 0 | Girls | Preschool |
| 8 | 1.415 | 142.240 | 6 | 0 | 0 | Girls | Preschool |
| 9 | 1.987 | 148.590 | 8 | 0 | 0 | Girls | Primary |
| 10 | 1.942 | 152.400 | 9 | 0 | 0 | Girls | Primary |

# 虛擬變項

- 在進行迴歸分析時， 我們利用虛擬變項(dummy variables)來代表 agegp

| Subjects | age | agegp | agegp1 | agegp2 | agegp3 |
|---|---|---|---|---|---|
| 1 | 6 | Preschool | 1 | 0 | 0 |
| 2 | 7 | Primary | 0 | 1 | 0 |
| 3 | 12 | Primary | 0 | 1 | 0 |
| 4 | 14 | Secondary | 0 | 0 | 1 |
| 5 | 18 | Secondary | 0 | 0 | 1 |

# ANOVA和迴歸分析的關係

- 在進行迴歸分析時，我們利用兩個虛擬變項(dummy variables) 來代表年齡類別解釋變項
- 模型可以寫成：

$$fev = b_0 + b_1 agegp_2 + b_2 agegp_3 + e$$

```
# R code:
lm2<-lm(fev~agegp, data=fev)
summary(lm2)
```

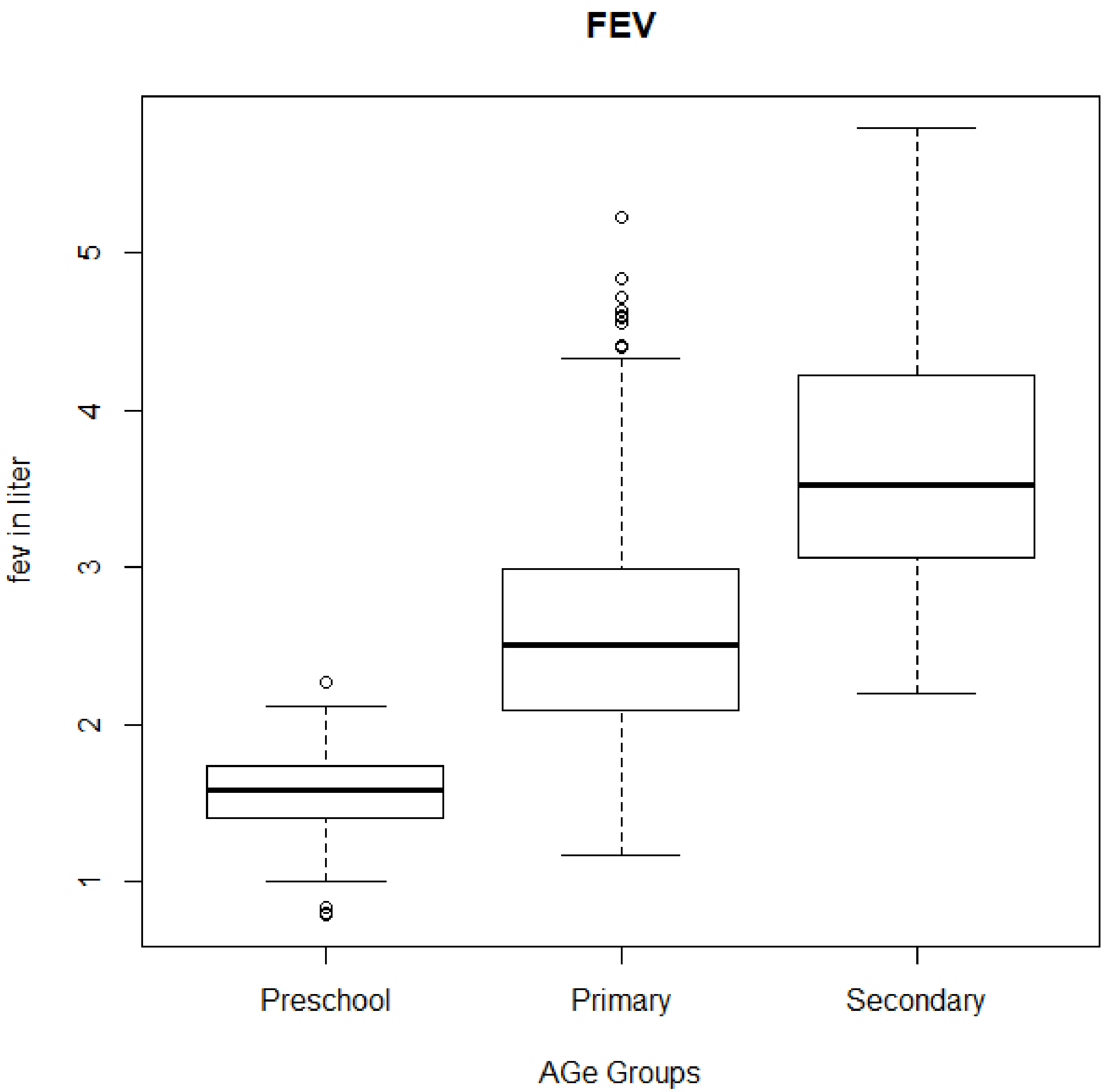
# ANOVA和迴歸分析的關係

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.56274    0.07692   20.32   <2e-16 ***
agegpPrimary    1.00679    0.08302   12.13   <2e-16 ***
agegpSecondary  2.03669    0.09879   20.62   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6706 on 651 degrees of freedom
Multiple R-squared:  0.4037,    Adjusted R-squared:  0.4019
F-statistic: 220.4 on 2 and 651 DF,  p-value: < 2.2e-16
```

K－1所估計是差異，不是平均值

# ANOVA和迴歸分析的關係



# ANOVA和迴歸分析的關係

```
> age.aov<-aov(fev~agegp,data=fev)
> summary(age.aov)
             Df Sum Sq Mean Sq F value Pr(>F)
agegp         2  198.2   99.09   220.4 <2e-16 ***
Residuals   651  292.7    0.45
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

- t-test 和 ANOVA 都算是 General Linear Model

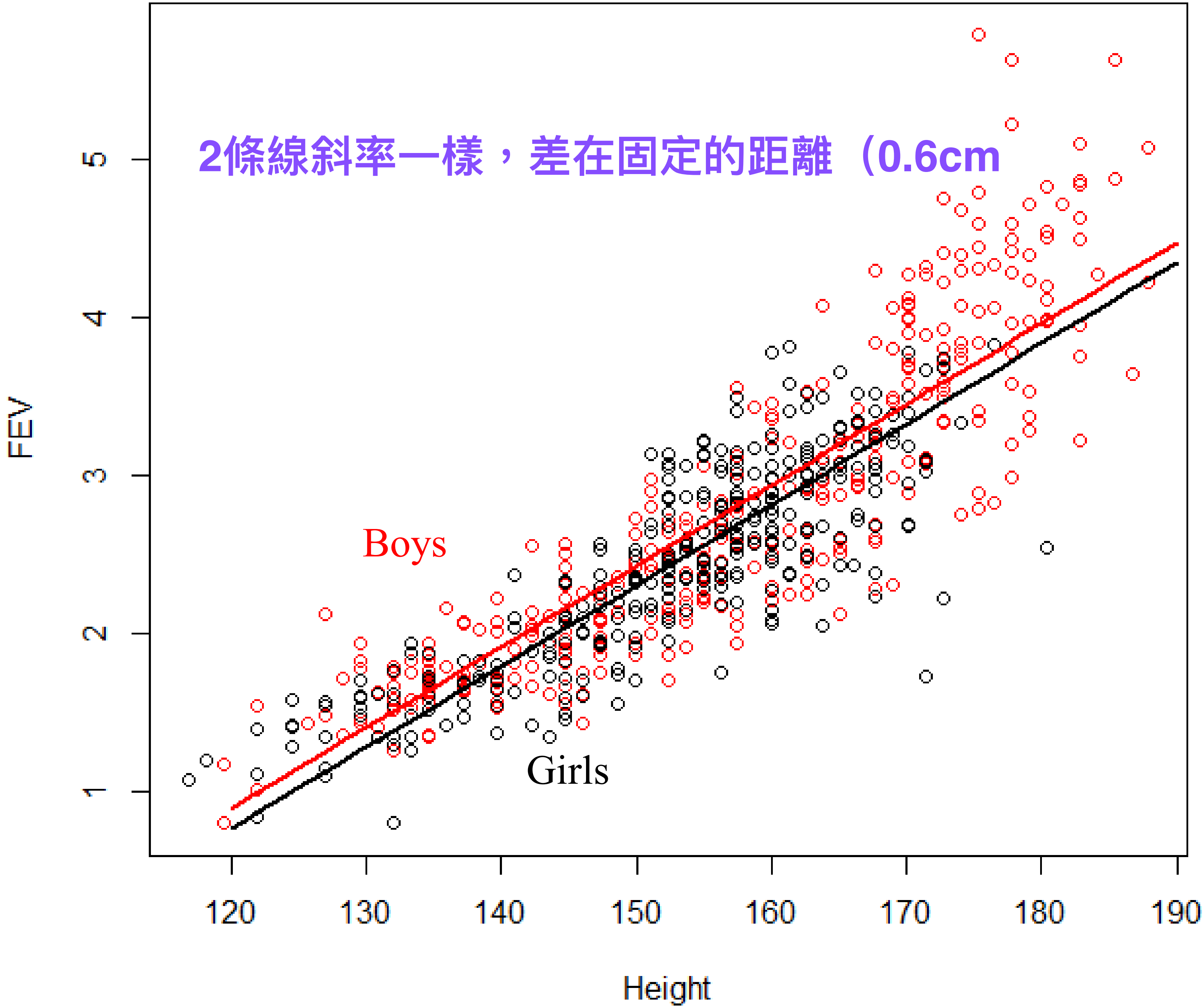
# 一個類別和一個連續解釋變項 共變因素分析

- 假設現在我們想知道身高(height) 、性別(gender)和最大呼氣量(fev)的關係，其中男生設為 1 ，女生設為 0：

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.390263   0.180082 -29.932  < 2e-16 ***
height       0.051272   0.001167  43.933  < 2e-16 ***
gender       0.125123   0.033801   3.702 0.000232 ***
```

$$fev = -5.39 + 0.051height + 0.125gender + e$$

當身高為0時，只要是男生身高就比女生多0.6公分


2條線斜率一樣，差在固定的距離（0.6cm
Boys
Girls
FEV
Height
1
2
3
4
5
120
130
140
150
160
170
180
190

這等於我們有兩個迴歸方程式，一個是針對男生，另一個針對女生：

男生：

$$\widehat{fev} = -5.39 + 0.051height + 0.125$$

女生：

$$\widehat{fev} = -5.39 + 0.051height$$

兩者的差別是一個常數 0.125，反映在兩條直線的截距不同。**也就是說平均而言， 男生的最大呼氣量比同身高女生大 0.125 公升**。

# 複迴歸

- 現在我們把身高、年齡、和性別都放到迴歸分析模型裏：

$$fev = b_0 + b_1 height + b_2 age + b_3 gender$$

- 結果如下：

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.448560   0.222966 -19.952  < 2e-16
height       0.041165   0.001872  21.986  < 2e-16
age          0.061364   0.009069   6.766 2.96e-11
gender       0.161112   0.033125   4.864 1.45e-06

Residual standard error: 0.4126 on 650 degrees of freedom
Multiple R-squared:  0.7746,    Adjusted R-squared:  0.7736
```

# 複迴歸

$$fev = -4.45 + 0.04height + 0.06age + 0.16gender$$

- 所以一個 15 歲、身高 170 公分的男孩，根據上面的模型，他的 fev 預測值為：

$$3.41 = -4.45 + 0.04 * 170 + 0.06 * 15 + 0.16$$

- 而一個 15 歲、身高 170 公分的女孩，根據上面的模型，她的 fev 預測值為：

$$3.25 = -4.45 + 0.04 * 170 + 0.06 * 15$$

# Assumptions for Multiple Linear Regression

- 解釋變數和反應變數之間的關係是直線的(linear)
  - 可以檢查殘差值對解釋變數的散佈圖 (scatterplot)
- 殘差值在解釋變數的任何一值之下都是服從常態分佈
  - 可以檢查殘差值對解釋變數的散佈圖或殘差值的 Q-Q plot

# Assumptions for Multiple Linear Regression

- 殘差值在解釋變數的任何一值之下的變異數(variance)都是相同的。這叫做**變異數恆定**假設(“constant variance” or “homoscedasticity”)
  - 可以檢查殘差值對解釋變數或適配值(fitted values)的散佈圖
- 殘差值是各別獨立的

# Height vs Residuals