Linear Regression Analysis (4): Polynomial Regression & Interaction

★

考分析的結果判讀
中心化對迴歸的影響
一次分有什麼意義

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

項式迴歸模型的使用

• 多項式迴歸模型兩個基本的使用型態為:

1.真正的曲線反應函數的確為一個多項式函數。

2.真正的曲線反應函數未知或是非常複雜,但是透過多項式迴歸模型可以做出不錯的近似效果。• 採用多項式迴歸模型應注意的是外插風險,尤其是在高 階的多項式上,雖然多項式迴歸模型對手上的資料做出 不錯的配適,不過一旦在資料範圍外進行外插時,很有 可能會轉移到預期外的方向上。

Polynomial Regression

「多項式迴歸」。迴歸函數採用多項式函數。誤差採用平方誤差。

演算法仍是 Normal Equation 。

1. 使用時機: 以自變項的多項式預測一個因變項。

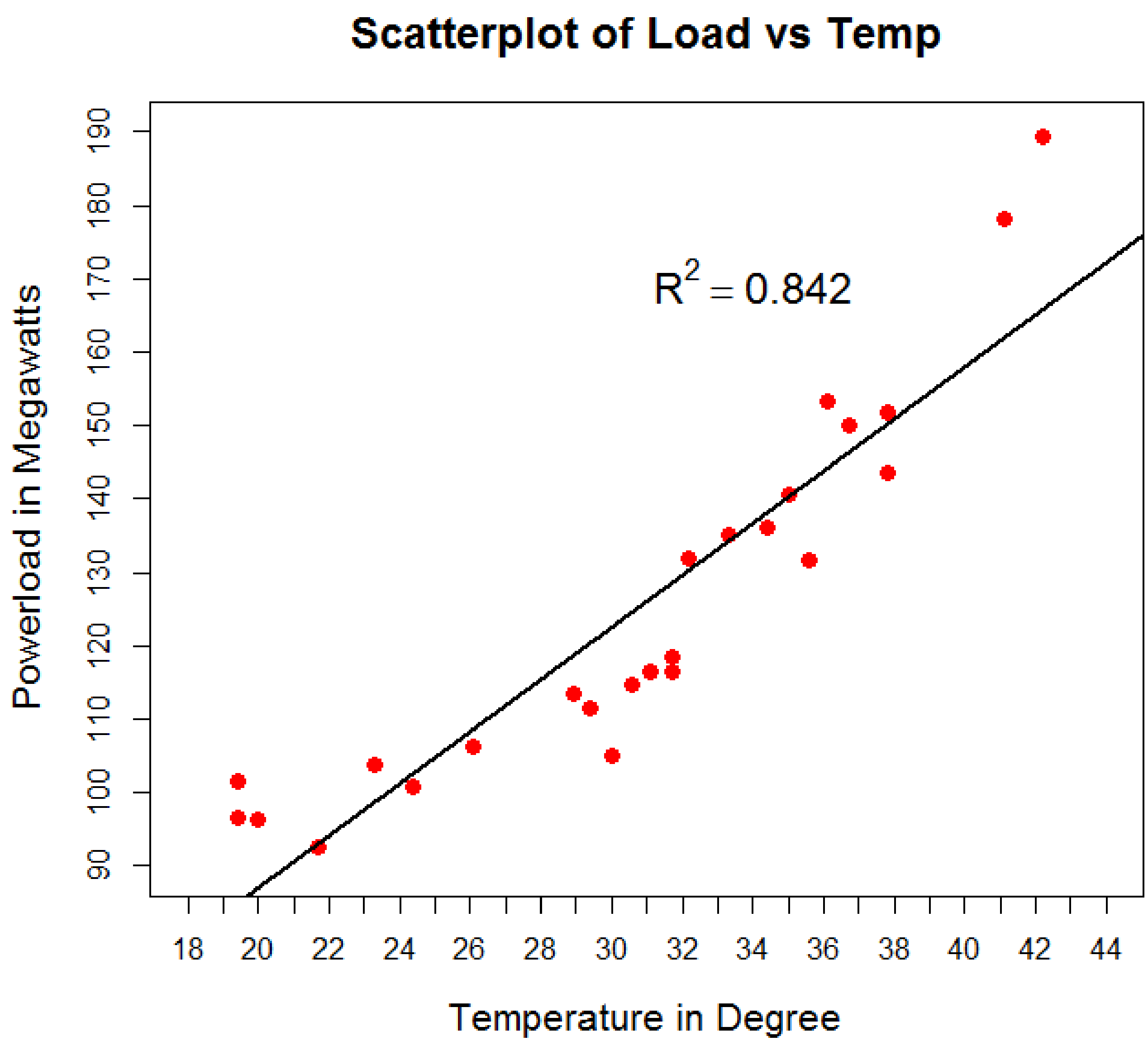
2. 分析類型: 回歸分析(regression analysis)。

Case Study: Peak Power Load & Temperature

- To operate efficiently, power companies must be able to predict the peak power load at their various stations.
- Peak power load is the maximum amount of power that must be generated each day to meet demand.
- A power company wants to use daily high temperature, x, to model daily peak power load, y, during the summer months when demand is greatest.
- Although the company expects peak load to increase as the temperature increases, the *rate* of increase in $E(y)$ might not remain constant as x increases.

- For example, a 1-unit increase in high temperature from 36°C to 37°C might result in a larger increase in power demand than would a 1-unit increase from 26°C to 27°C.
- Therefore, the company postulates that the relationship between temperature and power load may not be linear.
- A random sample of 25 summer days is selected and both the peak load (measured in megawatts) and high temperature (in Celsius degrees) recorded for each day.

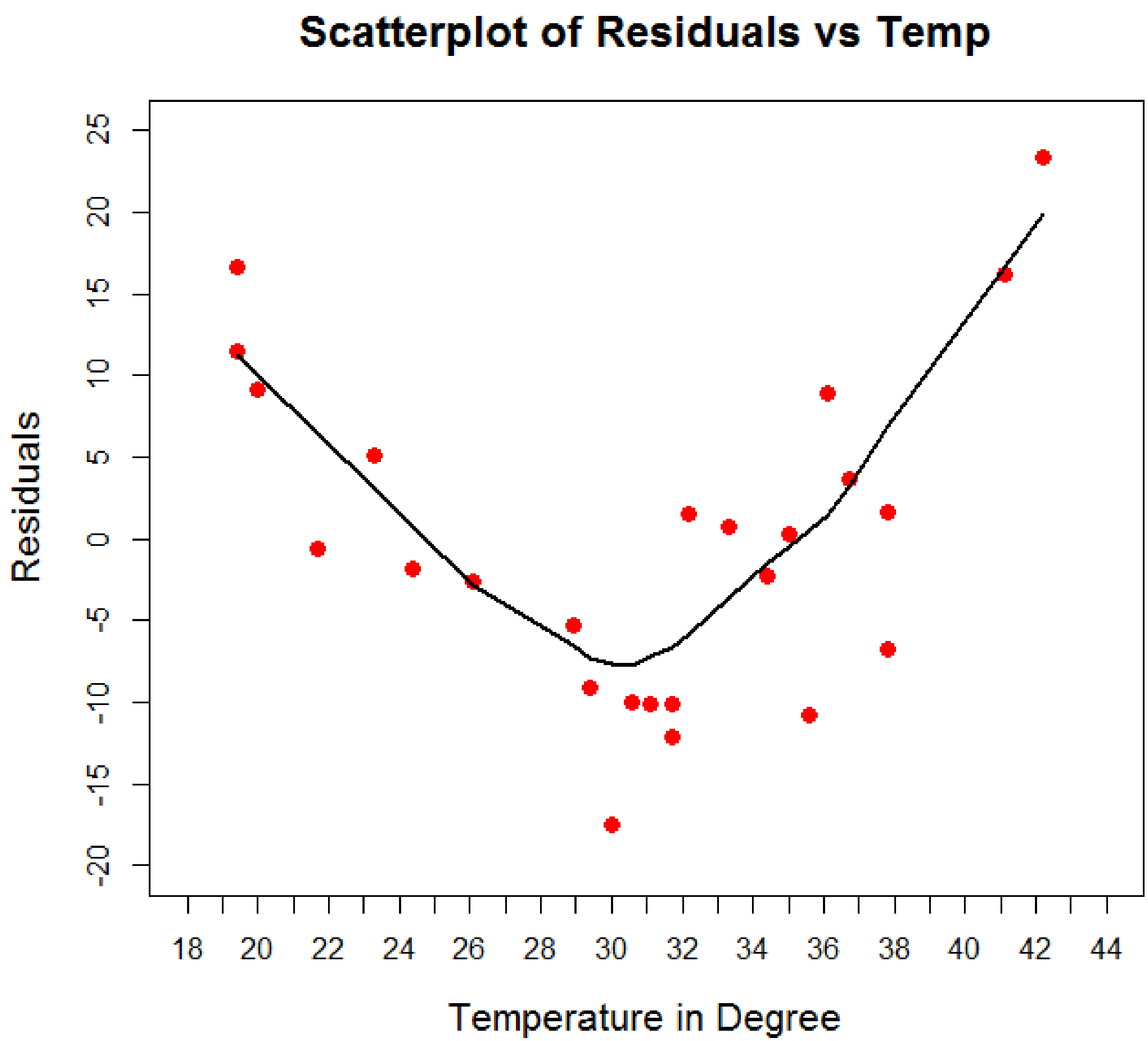
Linear Regression Model



Load	Celsius
136	34.4
131.7	35.6
140.7	35.0
189.3	42.2
96.5	19.4
116.4	31.1
118.5	31.7
113.4	28.9
132	32.2
178.2	41.1
101.6	19.4
92.5	21.7
151.9	37.8
106.2	26.1
153.2	36.1
150.1	36.7
114.7	30.6
100.9	24.4
96.3	20.0
135.1	33.3
143.6	37.8
111.4	29.4
116.5	31.7
103.9	23.3
105.1	30.0

```
> lm1<-lm(Load~Celsius, data=powerload)
> summary(lm1)

Residuals:
     Min       1Q   Median       3Q      Max 
-17.5024  -9.0725  -0.6394   5.0810  23.3906 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  16.1096    10.0946   1.596    0.124    
Celsius       3.5498     0.3208  11.066 1.09e-10 ***

Residual standard error: 10.37 on 23 degrees of freedom
Multiple R-squared:  0.8419,    Adjusted R-squared:  0.835 
F-statistic: 122.4 on 1 and 23 DF,  p-value: 1.095e-10
```

Residual Plot



- It is quite obvious the relationship between power load and temperature is not linear
- There are several ways to estimate a curvilinear or non-linear relationship, such as polynomials, spline and nonlinear functions.
- Polynomial regression is the most widely used method to model a non-linear relationship between an explanatory variable and the outcome

Polynomial Regression

- First-order model with one covariate: $\hat{y} = b_0 + b_1 x_1$
- Second-order (quadratic) model with one covariate: $\hat{y} = b_0 + b_1 x_1 + b_2 x_1^2$
- Third-order (cubic) model with one covariate: $\hat{y} = b_0 + b_1 x_1 + b_2 x_1^2 + b_3 x_1^3$
- *P*th-order model with one covariate: $\hat{y} = b_0 + b_1 x_1 + b_2 x_1^2 + \cdots + b_p x_1^p$
- Although the fitted line is not a straight line, these models are still "linear" model, usually known as curvilinear models to be distinguished from non-linear models

呈現複雜的非線性關係

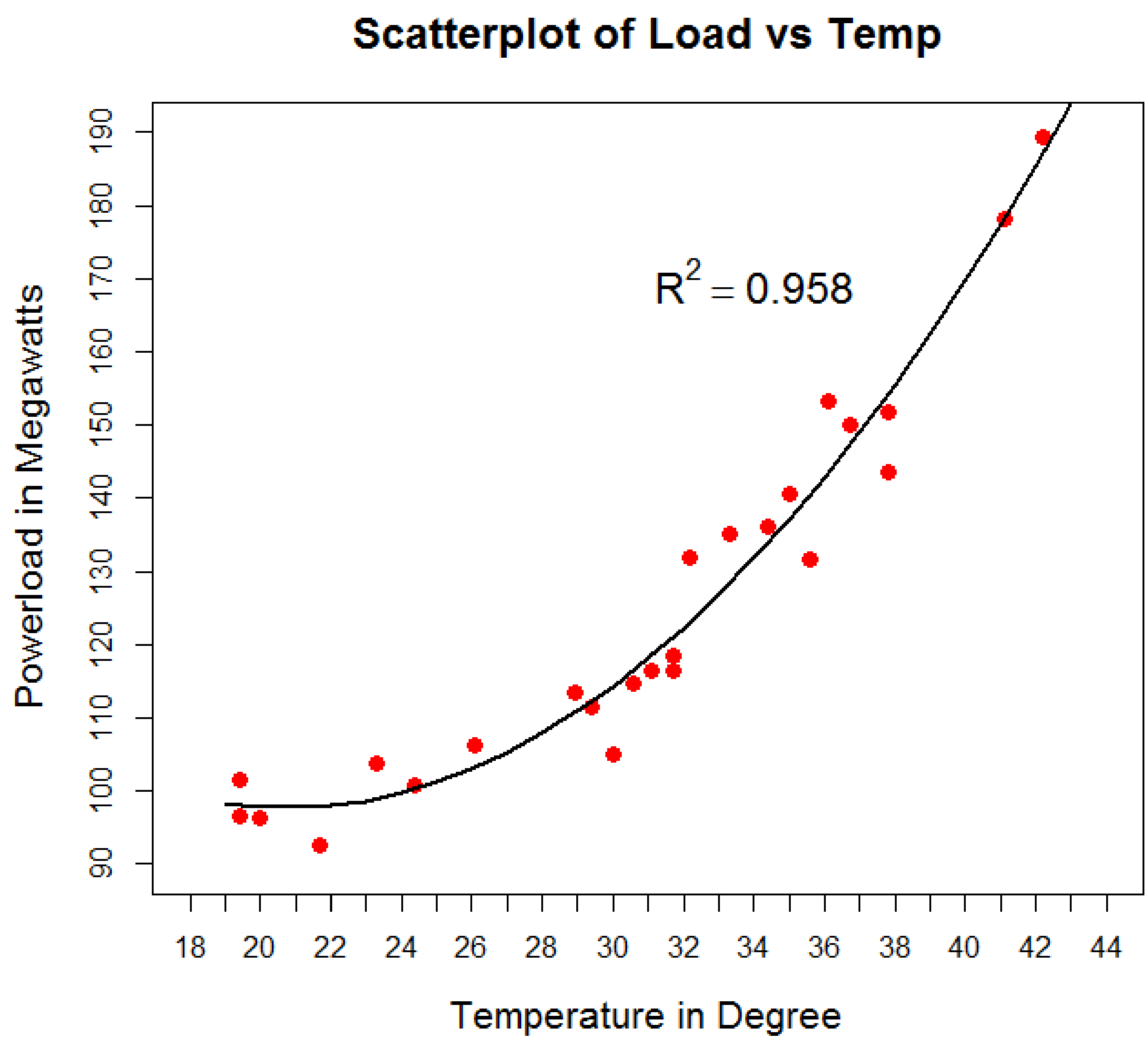
- To fit a quadratic model for our case study, we first create a new variable Celsius2, which is squared Celsius

```
> Celsius2 <- powerload$Celsius^2
> lm2<-lm(Load~Celsius+Celsius2, data=powerload)
> summary(lm2)
Residuals:
     Min        1Q    Median        3Q       Max
-10.5597   -2.2597    0.0827    2.9870    9.7328
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 181.17159   21.71016   8.345 2.91e-08 ***
Celsius      -8.04877    1.48894  -5.406 1.98e-05 ***
Celsius2      0.19403    0.02475   7.840 8.24e-08 ***

Residual standard error: 5.445 on 22 degrees of freedom
Multiple R-squared:  0.9583,    Adjusted R-squared:  0.9545
F-statistic: 252.9 on 2 and 22 DF,  p-value: 6.595e-16
```

當溫度為0時他的切線斜率－8

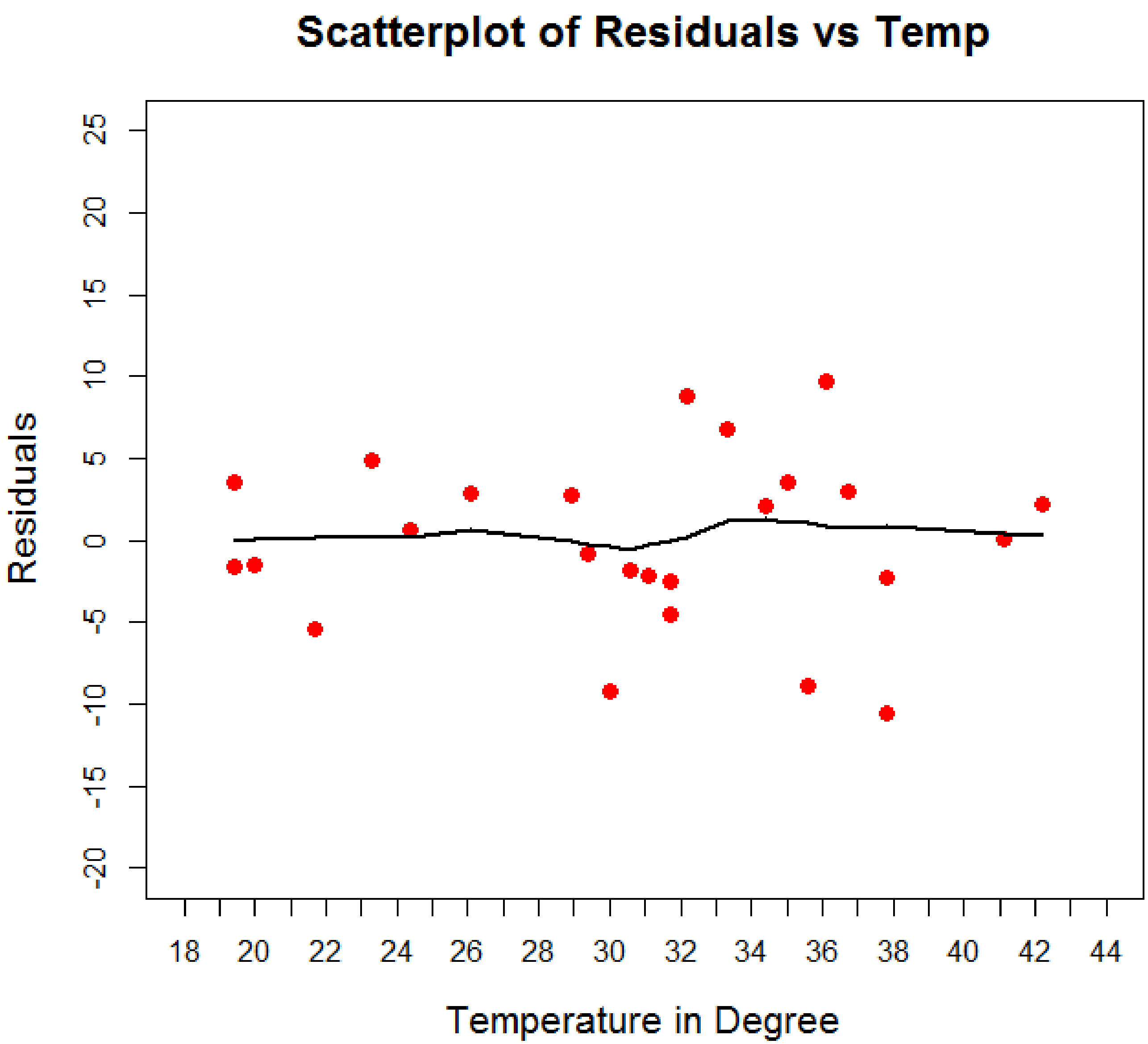
每上升1度

溫度的平方系數


Scatterplot of Load vs Temp
Powerload in Megawatts
90
100
110
120
130
140
150
160
170
180
190
$R^2 = 0.958$
18
20
22
24
26
28
30
32
34
36
38
40
42
44
Temperature in Degree

- Both Celsius and Celsius2 are highly significant
- The overall model is also significant. This is reflected by the increased R^2 (from 0.842 to 0.958) and the significant F-test result (F = 252.9, df = (2, 22))

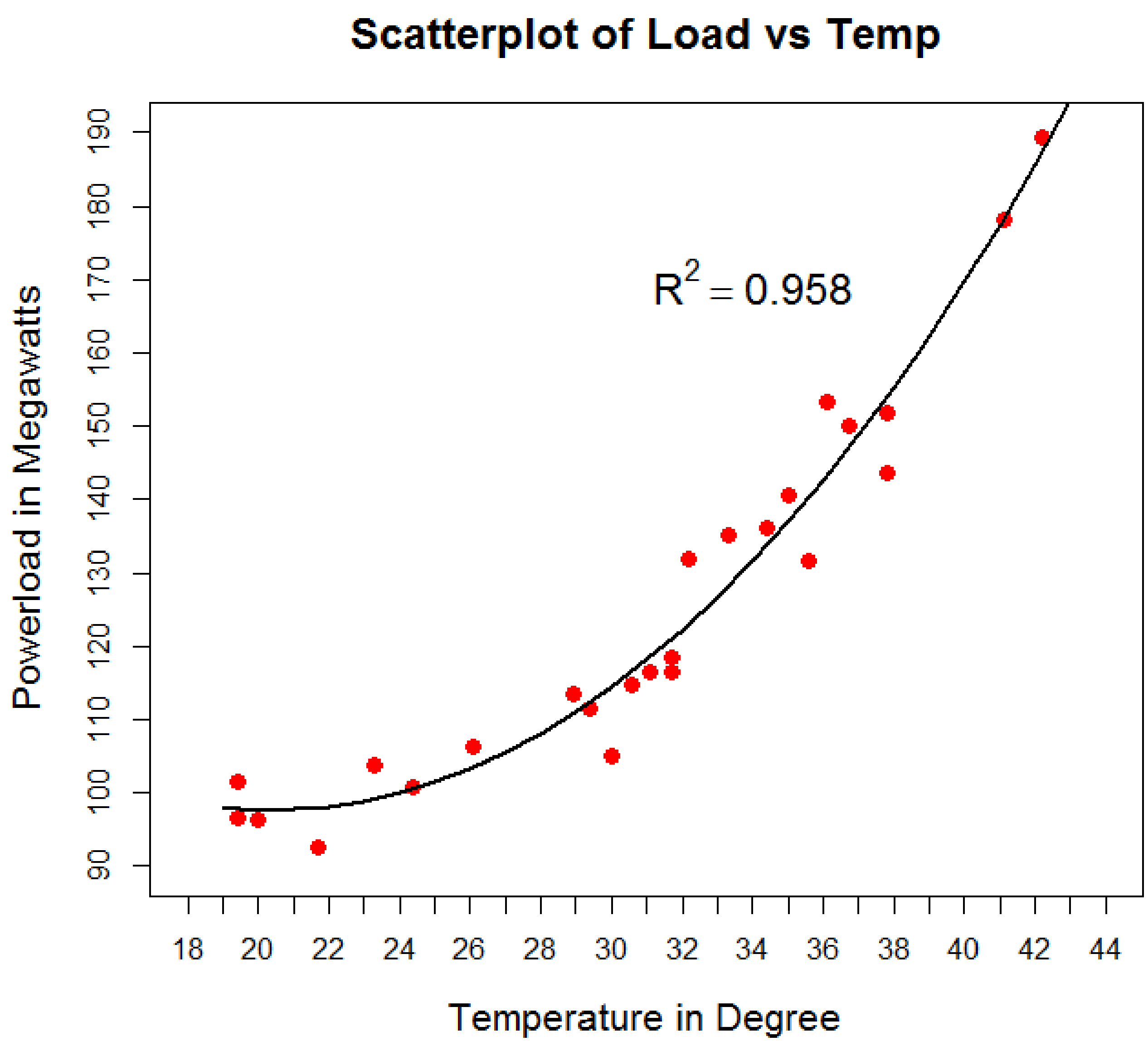
Residual Plot



將溫度三次方後，結果P值都不顯著
但R－squared：0.9584
最有可能因素為變數共線性高

- Now, let us see if a cubic model will further improve the model fit:

```
> Celsius3 <- powerload$Celsius^3
> lm3<-lm(Load~Celsius+Celsius2+Celsius3, data=powerload)
> summary(lm3)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.645e+02  1.159e+02   1.419    0.171
Celsius     -6.293e+00  1.205e+01  -0.522    0.607
Celsius2     1.349e-01  4.036e-01   0.334    0.742
Celsius3     6.429e-04  4.378e-03   0.147    0.885

Residual standard error: 5.57 on 21 degrees of freedom
Multiple R-squared:  0.9584,    Adjusted R-squared:  0.9524
F-statistic: 161.1 on 3 and 21 DF,  p-value: 1.186e-14
```


Scatterplot of Load vs Temp
$R^2 = 0.958$
Powerload in Megawatts
Temperature in Degree
90
100
110
120
130
140
150
160
170
180
190
18
20
22
24
26
28
30
32
34
36
38
40
42
44

多0.0001,顯著R值變高，但變數之間有高度共線性，故P值不顯著

- The model R^2 only increases marginally from 0.9583 to 0.9584.
- However, none of the explanatory variables is statistically significant! But the *F*-test remains highly significant
- So the model is good but none of the explanatory variables make “important” contribution. How did this happen?
- This is because the linear, quadratic and cubic terms are highly collinear (around 0.97)!
- But why did collinearity not cause any problem for quadratic model?

Centering

中心化，只能處理多向性迴歸所引起的共線性，影響最高係數
影響截距而已，影響小一個的系數和截距，
不一定要中心化，也不常做，可以用多向式去算，唯一的好處是好解釋
把溫度中心化後再去分析，再放心的保留一二次方向，第三次可去除

- Centering is useful for reducing the collinearity between a variable and its power terms
- Note that centering does not work for other collinearities

```
> ## centering Celsius
> Celsius.c <- powerload$Celsius -
mean(powerload$Celsius)
> Celsius.c2 <- Celsius.c^2
> Celsius.c3 <- Celsius.c^3
> lm4<-lm(Load~Celsius.c+Celsius.c2+Celsius.c3,
data=powerload)
```

- Note that the mean of Celsius is 30.8

描述加速度過程中的變化，第三次方向是不是顯著

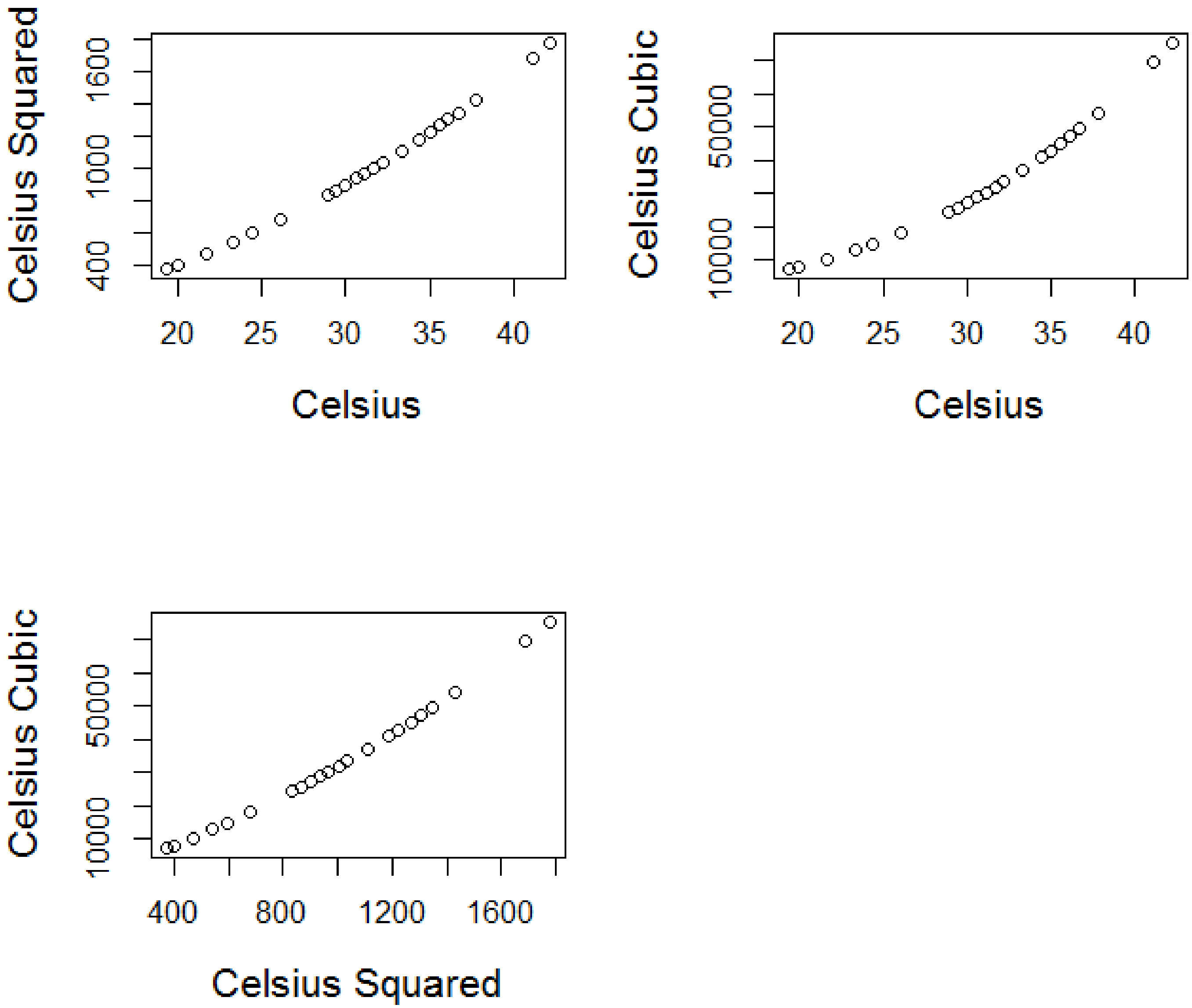
```
> summary(lm4)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.4228  -2.1391  -0.0845   3.1520   9.9008 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.174e+02  1.559e+00  75.280  < 2e-16 ***
Celsius.c   3.843e+00  4.353e-01   8.829 1.64e-08 ***
Celsius.c2  1.943e-01  2.537e-02   7.656 1.65e-07 ***
Celsius.c3  6.429e-04  4.378e-03   0.147    0.885    

Residual standard error: 5.57 on 21 degrees of freedom
Multiple R-squared:  0.9584,    Adjusted R-squared:  0.9524 
F-statistic: 161.1 on 3 and 21 DF,  p-value: 1.186e-14
```

How Does Centering Work?

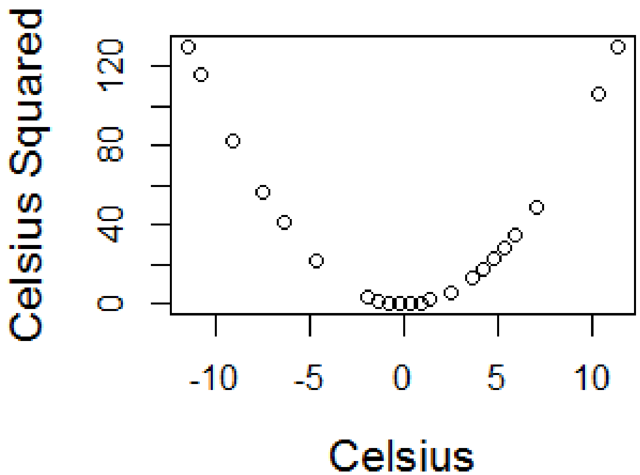
- After centering, the model correctly shows that both the linear and quadratic terms are statistically significant, while the cubic term is not.
- Note that centered model has the same R^2 as the original model. These two models are identical.

並不影響畫出的線和模型，但會影響一次二次三次方的關係

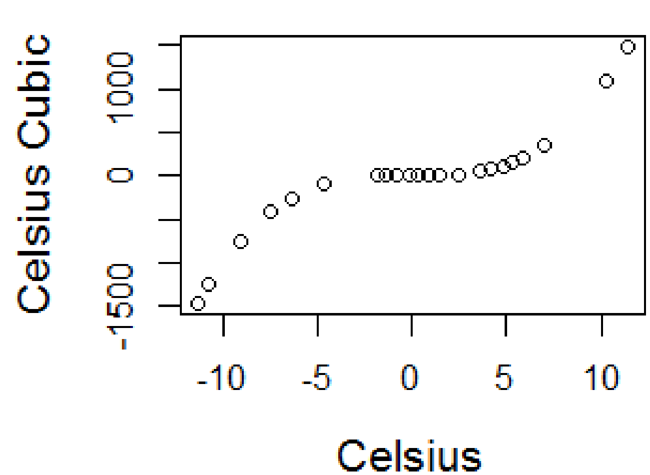
Correlations between Celsius & Its Power Terms



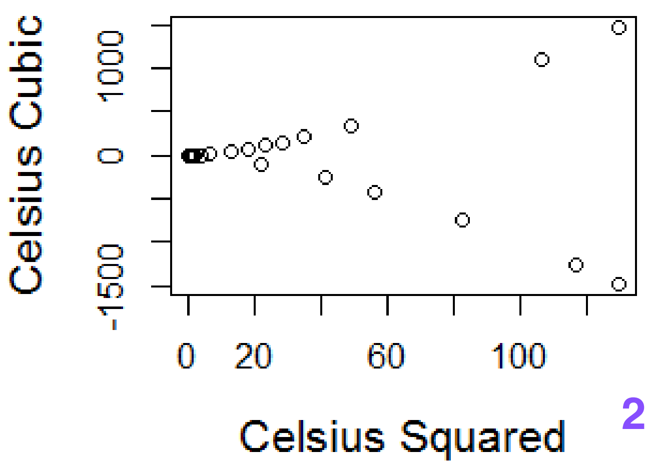
Correlations between Celsius & Its Power Terms After Centering



中心化後，



文字



2,3無關

Impact of Centering

We now use the quadratic model to illustrate the impact of centering on polynomial regression model.

Recall the original model:

```
> lm2<-lm(Load~Celsius+Celsius2, data=powerload)
> summary(lm2)
```

當溫度為0時，切線的斜率為______

```
Estimate Std. Error t value Pr(>|t|)
(Intercept) 181.17159   21.71016   8.345 2.91e-08 ***
Celsius      -8.04877    1.48894  -5.406 1.98e-05 ***
Celsius2      0.19403    0.02475   7.840 8.24e-08 ***

Residual standard error: 5.445 on 22 degrees of freedom
Multiple R-squared:  0.9583,   Adjusted R-squared:  0.9545
F-statistic: 252.9 on 2 and 22 DF,  p-value: 6.595e-16
```

Impact of Centering

中心化後

會讓座標平移，但圖和模型不變

We now re-fit the model using the centered Celsius:

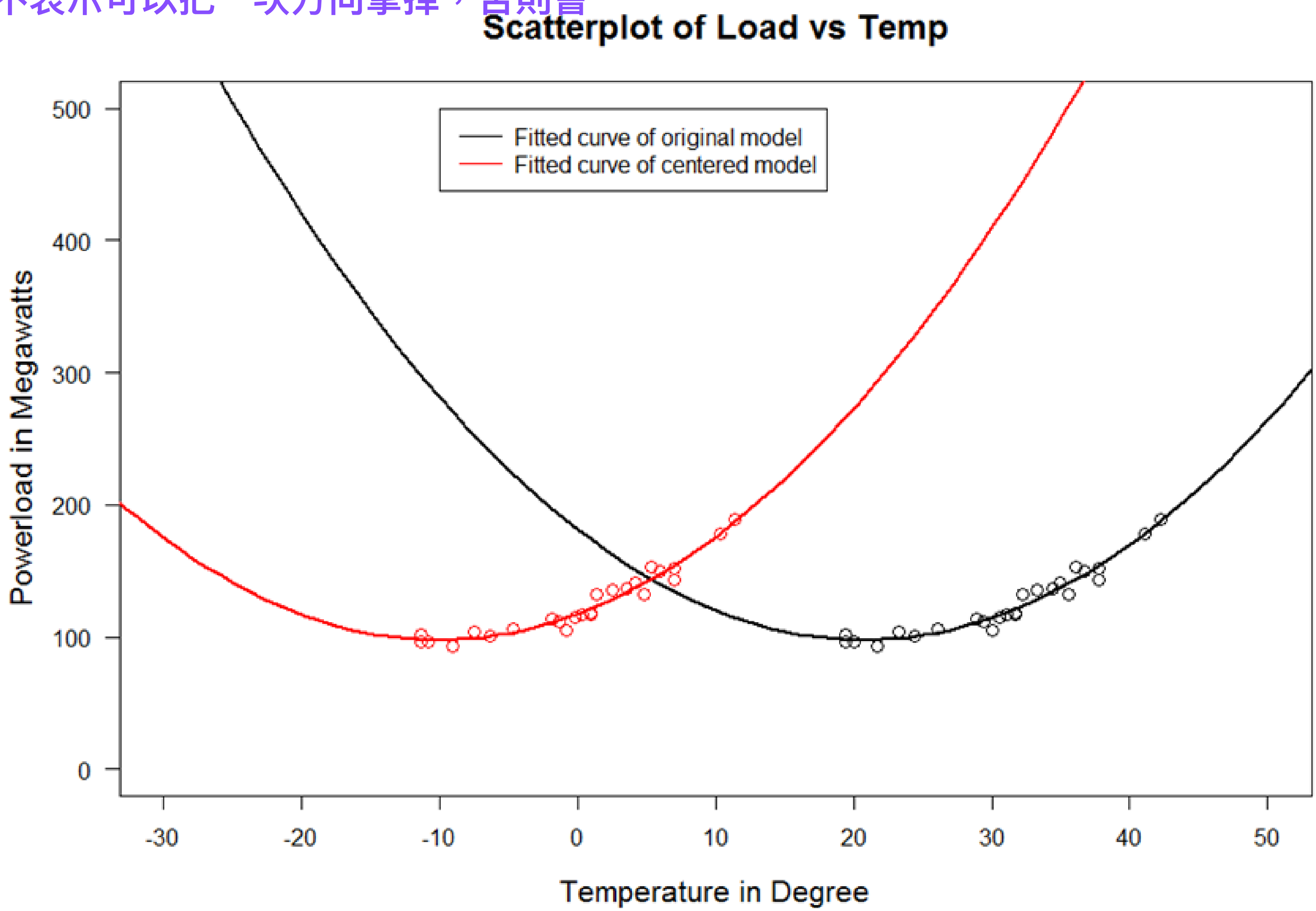
```
> # fit the centered quadratic model
> lm5<-lm(Load~Celsius.c+Celsius.c2, data=powerload)
> summary(lm5)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 117.31439    1.50227   78.09  < 2e-16 ***
Celsius.c     3.90166    0.17427   22.39  < 2e-16 ***
Celsius.c2    0.19403    0.02475    7.84 8.24e-08 ***
---
Residual standard error: 5.445 on 22 degrees of freedom
Multiple R-squared:  0.9583,    Adjusted R-squared: 
0.9545
F-statistic: 252.9 on 2 and 22 DF,  p-value: 6.595e-16
```

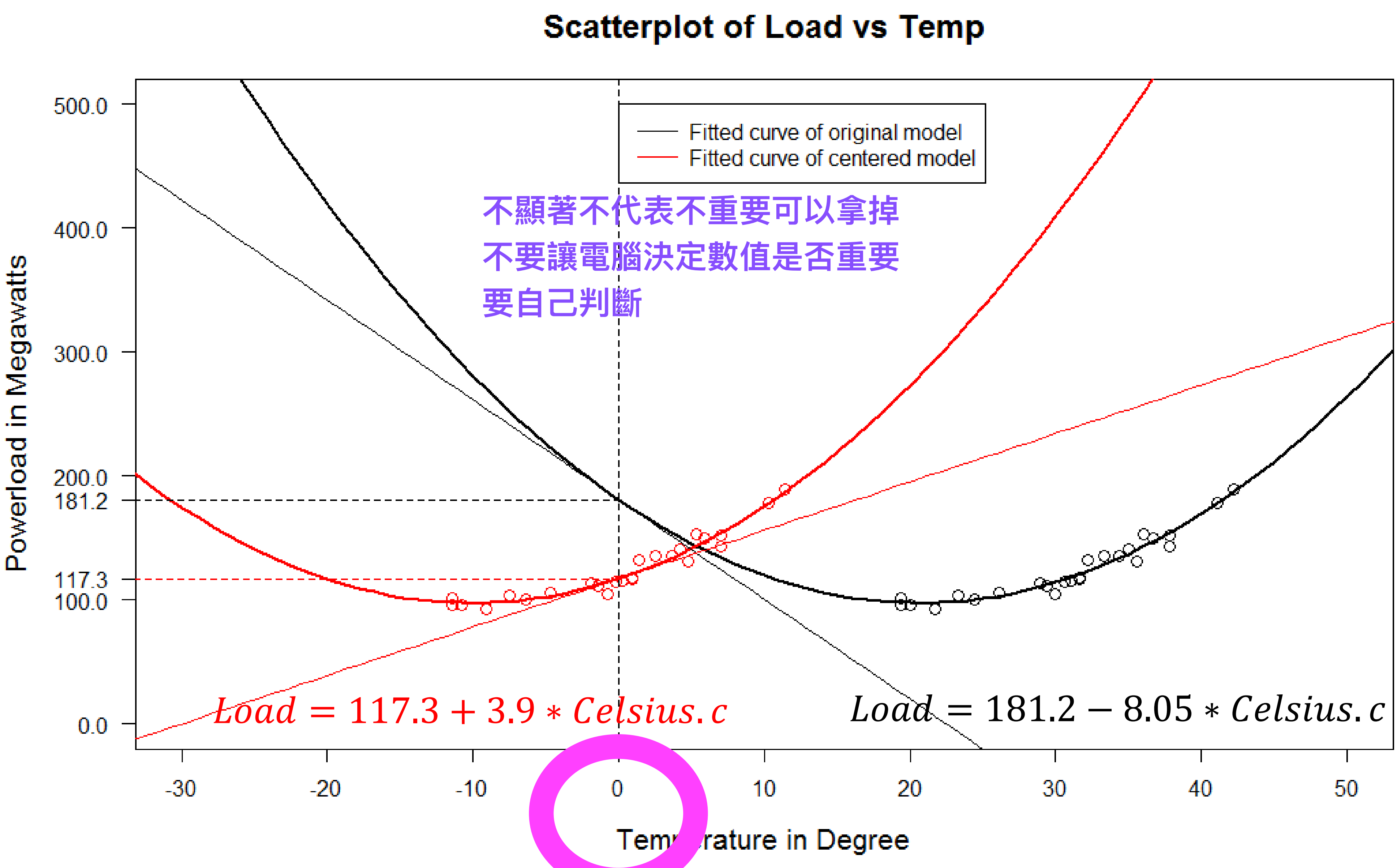
Impact of Centering

- Note that the regression coefficients for intercept and linear terms are different in the two models, while the coefficient for the quadratic term remain unchanged.
- Centering at mean value of temperature is equivalent to moving the fitted parabolic curve horizontally to the left by 30.8
- Because the shape of the curve remains unaffected, the coefficient for quadratic term remain unchanged

十人田土

座標平移時，截距不同了
一次方向係數也跟著改變
不表示可以把一次方向拿掉，否則會




Scatterplot of Load vs Temp
Fitted curve of original model
Fitted curve of centered model
不顯著不代表不重要可以拿掉
不要讓電腦決定數值是否重要
要自己判斷
Powerload in Megawatts
500.0
400.0
300.0
200.0
181.2
117.3
100.0
0.0
$Load = 117.3 + 3.9 * Celsius.c$
$Load = 181.2 - 8.05 * Celsius.c$
-30
-20
-10
0
10
20
30
40
50
Temperature in Degree

Interpretation of Polynomial Regression Coefficients

For the original polynomial model,

$$Load = 181.2 - 8.05Celsius + 0.194Celsius^2$$

- The coefficient for intercept (181.2) is the estimated power load in megawatts when temperature is at zero Celsius
- The coefficient for slope (−8.05) for Celsius is the slope of the tangent line for the fitted parabolic curve when temperature is at zero Celsius

Interpretation of Polynomial Regression Coefficients

For the centered polynomial model,

$$Load = 117.3 + 3.90Celsius + 0.194Celsius^2$$

- The coefficient for intercept (117.3) is the estimated power load in megawatts when temperature is at 30.8 Celsius 當溫度瞬間為30.8時，用電量的斜率為3·9
- The coefficient slope (3.90) for Celsius is the slope of the tangent line for the fitted parabolic curve when temperature is at 30.8 Celsius

多向式要第一次第二次方向逐次加入，加到不顯著時即停止，不顯著時把高的次方拿掉，而不是低的冷方拿掉，會影響模式和斜率
常用在身高的

Interaction

類別變數間的交互作用

Interaction in Regression Analysis

一個類別變向（男或女），Y變向為身高，迴歸係數為男女身高平均值

二個類別變向（男和女），Y有四種結果，其迴歸係數為四種平均值的差異

- Interaction between two binary variables
  - Estimate means for four groups
- Interaction between one binary and one continuous variable
  - Estimate two regression lines with different slopes for the two groups
- Interaction between two continuous variables
  - Estimate a curved plane

Interaction between Two Binary Variables

- We use the FEV example of 654 children to illustrate:

Id	fev	age	gender	smoking	height
1	1.404	3	1	0	131
2	1.072	3	0	0	117
3	0.839	4	0	0	122
4	1.569	4	0	0	127
5	1.577	4	0	0	124
6	0.796	4	1	0	119
7	1.789	4	1	0	132
8	1.102	4	0	0	122
....	....	....	....	....	....
650	4.404	18	1	1	179
651	2.853	18	0	0	152
652	5.102	19	1	0	183
653	3.519	19	0	1	168
654	3.345	19	0	1	166

Interaction between Two Binary Variables

- We first create a new variable agecat which is coded "younger" for children <= 11 y/o and coded "older" for those > 11 y/o 分為小於11歲和大於的
- We then calculate the means for those children stratified by gender and agecat: 另一種為男和女，故有4組（變向）

```
FEV$agecat <- ifelse(FEV$age > 11, c("older"),
c("younger"))

FEV$sex <- ifelse(FEV$gender == 0, c("girls"),
c("boys"))

aggregate(x = FEV$fev, by = list(FEV$sex,
FEV$agecat), FUN = "mean")
```

Interaction between Two Binary Variables

```
 Group.1 Group.2        x
1    boys   older 3.933056
2   girls   older 2.986833
3    boys younger 2.402467
4   girls younger 2.258880
```

和2個類別變向（男女年紀小年紀大）

We now run a regression model with sex and agecat as covariates

```
lm1 <- lm(fev ~ sex + agecat, data = FEV)
summary(lm1)
```

文字

Interaction between Two Binary Variables

這個模型可以給出4組平均值，但和觀察到的可能不一樣

截距估計為3.6

性別控制年齡，

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.64862    0.05779  63.141  < 2e-16 ***
sex.girls      -0.35704    0.05338  -6.689 4.84e-11 ***
agecat.younger -1.14210    0.06037 -18.917  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 0.6823 on 651 degrees of freedom
Multiple R-squared:  0.3827,   Adjusted R-squared:  0.3809
F-statistic: 201.8 on 2 and 651 DF,  p-value: < 2.2e-16
```

The intercept 3.65 is the estimated mean fev for older boys, which is smaller than the real mean fev 3.93

Interaction between Two Binary Variables

- In this model, we assume the difference in mean fev between boys and girls is the same for younger and older children and vice versa,
- i.e. we also assume the difference in mean fev between younger and older children is the same in boys and girls
- However, it is very likely the difference in mean fev between boys and girls is greater for older children
- This is equivalent to an interaction between gender and age groups

年齡對FEV在不同姓別年齡層的影響
姓別對FEV

Interaction between Two Binary Variables

- To test the interaction, we create a new variable sex.age.i, which is product of agecat and sex

Id	fev	age	sex	smoking	height	agecat	sex.age.i
1	1.404	3	0 (boys)	0	131	1 (younger)	0
2	1.072	3	1 (girls)	0	117	1 (younger)	1
3	0.839	4	1 (girls)	0	122	1 (younger)	1
4	1.569	4	1 (girls)	0	127	1 (younger)	1
5	1.577	4	1 (girls)	0	124	1 (younger)	1
6	0.796	4	0 (boys)	0	119	1 (younger)	0
....	....	....	....	....	....	....	....
650	4.404	18	0 (boys)	1	179	0 (older)	0
651	2.853	18	1 (girls)	0	152	0 (older)	0
652	5.102	19	0 (boys)	0	183	0 (older)	0
653	3.519	19	1 (girls)	1	168	0 (older)	0
654	3.345	19	1 (girls)	1	166	0 (older)	0

Interaction between Two Binary Variables

```
> lm3 <- lm(fev ~ sex + agecat + sex.age.i, data = FEV)
> summary(lm3)
```

估計青少女：3.93－0.94

估計小女童3.93-0.94.1.53

```
Residuals:
     Min       1Q   Median       3Q      Max 
-2.01706 -0.50003 -0.00888  0.40919  2.23453 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      3.93306    0.06949  56.600  < 2e-16 ***
sex.girls       -0.94622    0.10001  -9.461  < 2e-16 ***
agecat.younger  -1.53059    0.08121 -18.847  < 2e-16 ***
sex.age.i        0.80264    0.11673   6.876 1.45e-11 ***
---
Residual standard error: 0.6592 on 650 degrees of freedom
Multiple R-squared:  0.4246,	Adjusted R-squared:  0.4219 
F-statistic: 159.9 on 3 and 650 DF,  p-value: < 2.2e-16
```

估計青少年

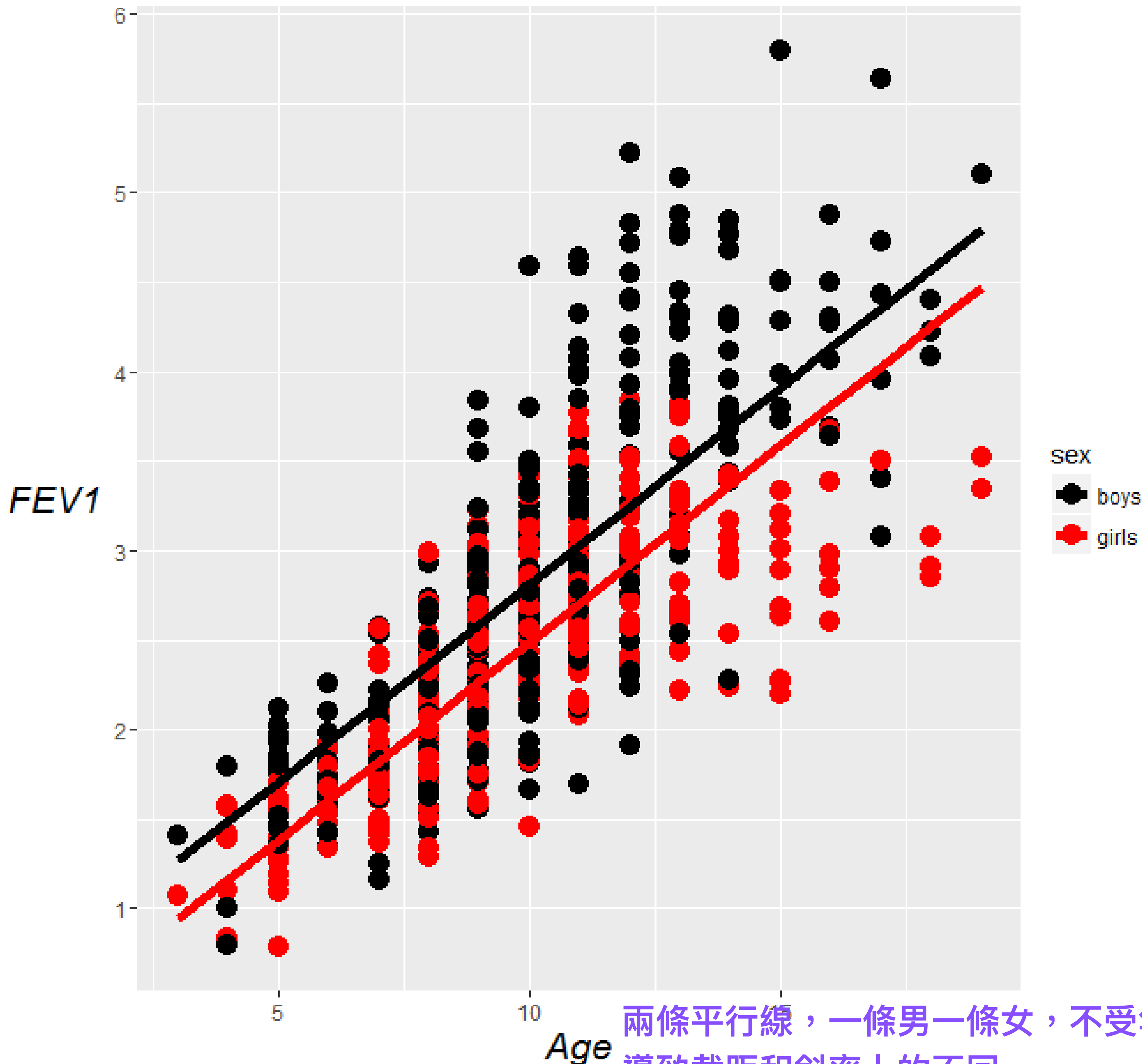
Interaction between Two Binary Variables

- The intercept 3.93 is the estimated mean fev for older boys, which is identical to the real mean fev
- We can work out the remaining means:
  - Older girls = 3.93 – 0.95 = 2.99
  - Younger boys = 3.93 – 1.53 = 2.40
  - Young girls = 3.93 – 0.95 – 1.53 + 0.80 = 2.26
- Those values are identical to their real means

Interaction between One Binary and One Continuous Variables

- Recall that in linear regression with one binary and one continuous covariates, the results are two fitted lines:

```
> lm4 <- lm(fev ~ sex + age, data = FEV)
> summary(lm4)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.604713   0.078124   7.740 3.79e-14 ***
sex.girls    -0.323335   0.042609  -7.588 1.13e-13 ***
age           0.220445   0.007215  30.553  < 2e-16 ***
```



兩條平行線，一條男一條女，不受年齡引響
導致截距和斜率上的不同

Interaction between One Binary and One Continuous Variables

- The interaction model is to fit two straight lines with different slopes for girls and boys

Id	fev	age	sex	smoking	height	agecat	sex.age
1	1.404	3	0 (boys)	0	131	1	0
2	1.072	3	1 (girls)	0	117	1	3
3	0.839	4	1 (girls)	0	122	1	4
4	1.569	4	1 (girls)	0	127	1	4
5	1.577	4	1 (girls)	0	124	1	4
6	0.796	4	0 (boys)	0	119	1	0
....	....	....	....	....	....	....	....
650	4.404	18	0 (boys)	1	179	0	0
651	2.853	18	1 (girls)	0	152	0	18
652	5.102	19	0 (boys)	0	183	0	0
653	3.519	19	1 (girls)	1	168	0	19
654	3.345	19	1 (girls)	1	166	0	19

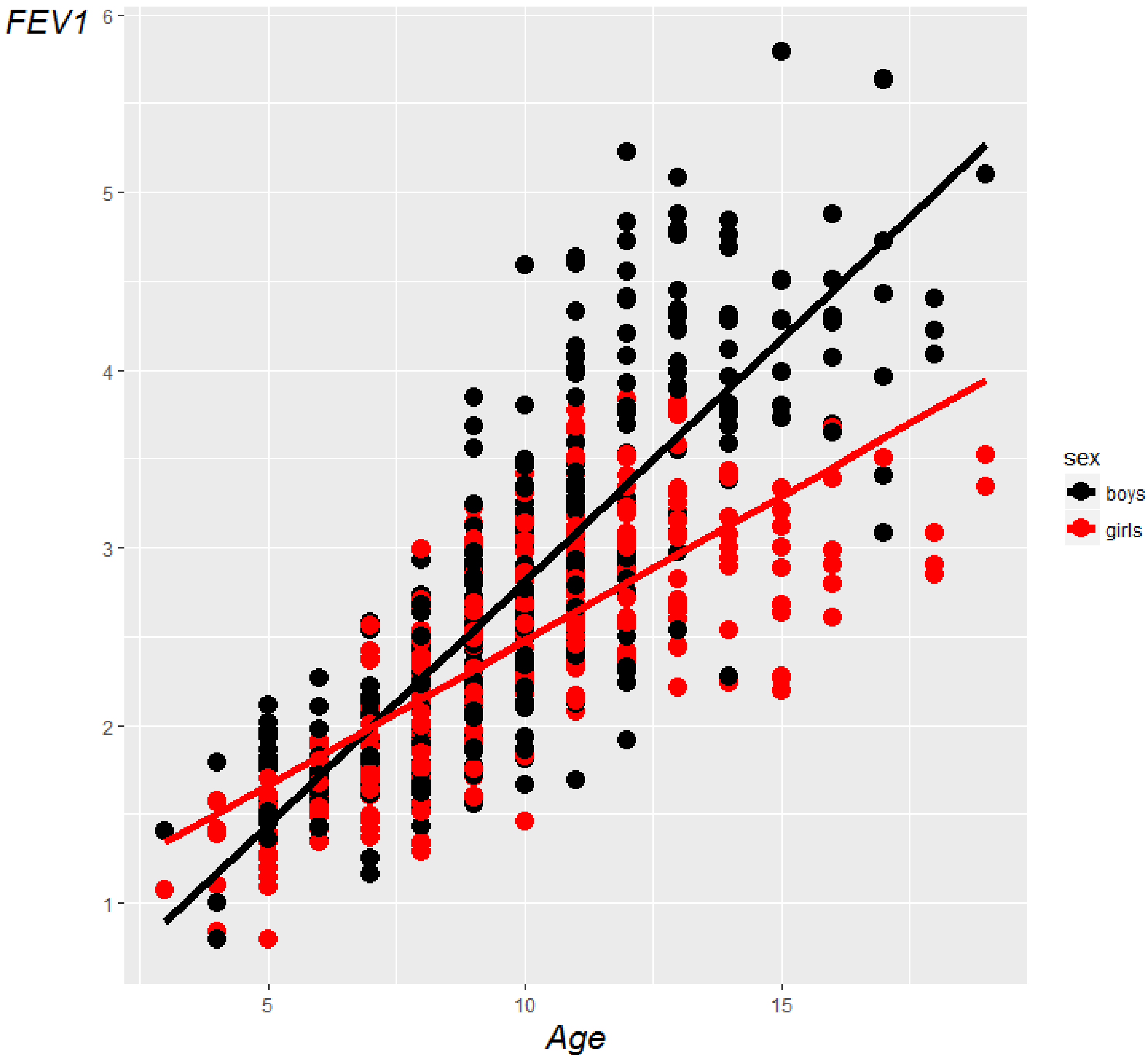
Interaction between One Binary and One Continuous Variables

```
> lm5 <- lm(fev ~ sex*age, data = FEV)
> summary(lm5)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.07360    0.09966   0.739     0.46
sex.girls         0.77587    0.14275   5.435 7.74e-08 ***
age               0.27348    0.00954  28.667  < 2e-16 ***
sex.girls:age    -0.11075    0.01379  -8.033 4.47e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 0.5196 on 650 degrees of freedom
Multiple R-squared:  0.6425,   Adjusted R-squared:  0.6408
F-statistic: 389.4 on 3 and 650 DF,  p-value: < 2.2e-16
```

考分析的結果判讀
中心化對迴歸的影響
一次分有什麼意義

文字

女孩的斜
率



Interaction Between Two Continuous Variables

- We now regress fev on both age and height:

$$\widehat{fev} = b_0 + b_1 age + b_2 height$$

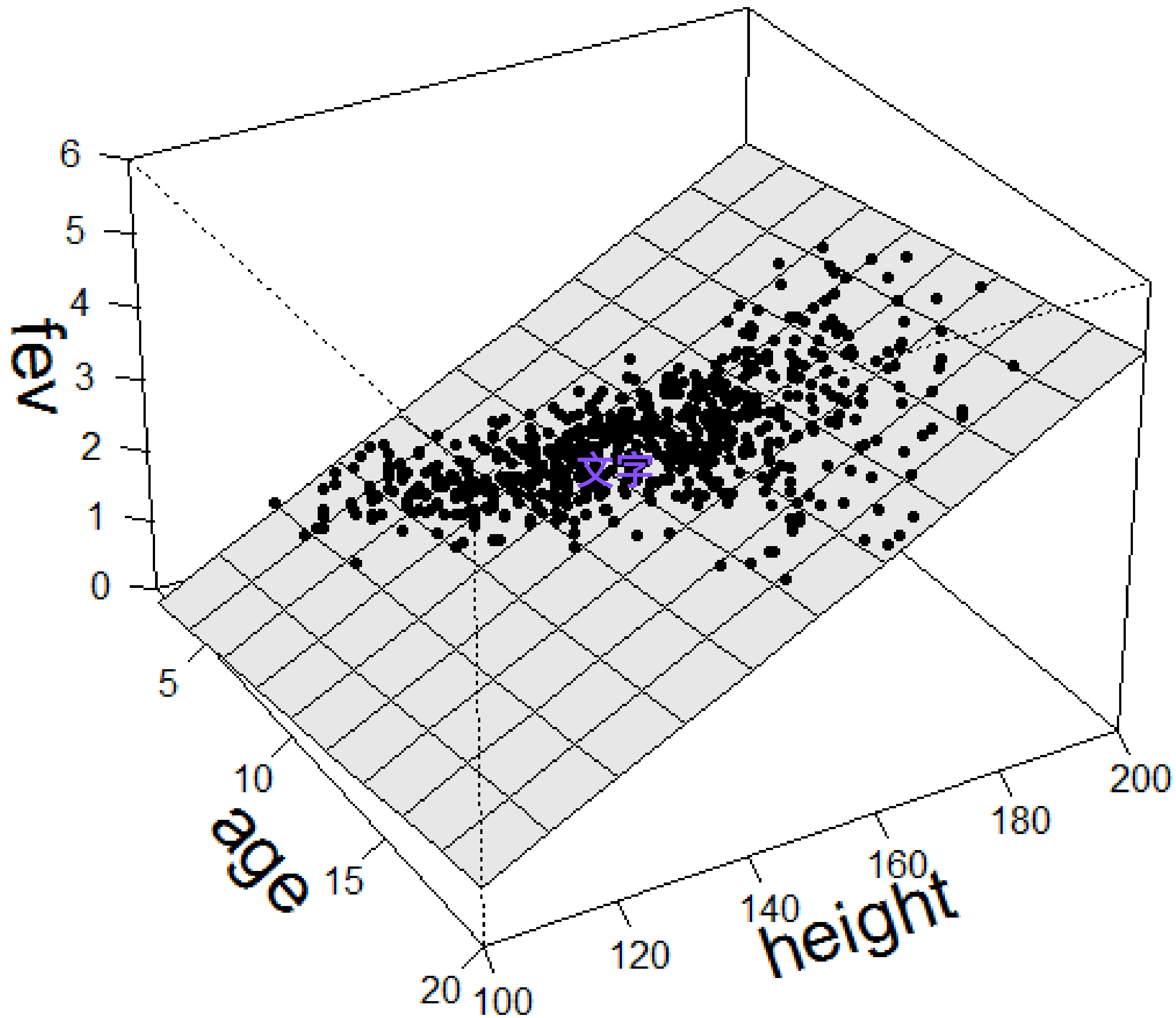
- The fitted values form a plane in a 3-dimensional space
- If we include an interaction between age and height, i.e. a product term into the model:

$$\widehat{fev} = b_0 + b_1 age + b_2 height + b_3 age * height$$

- The fitted valued form a curved plane

是平面的，

fev ~ age + height

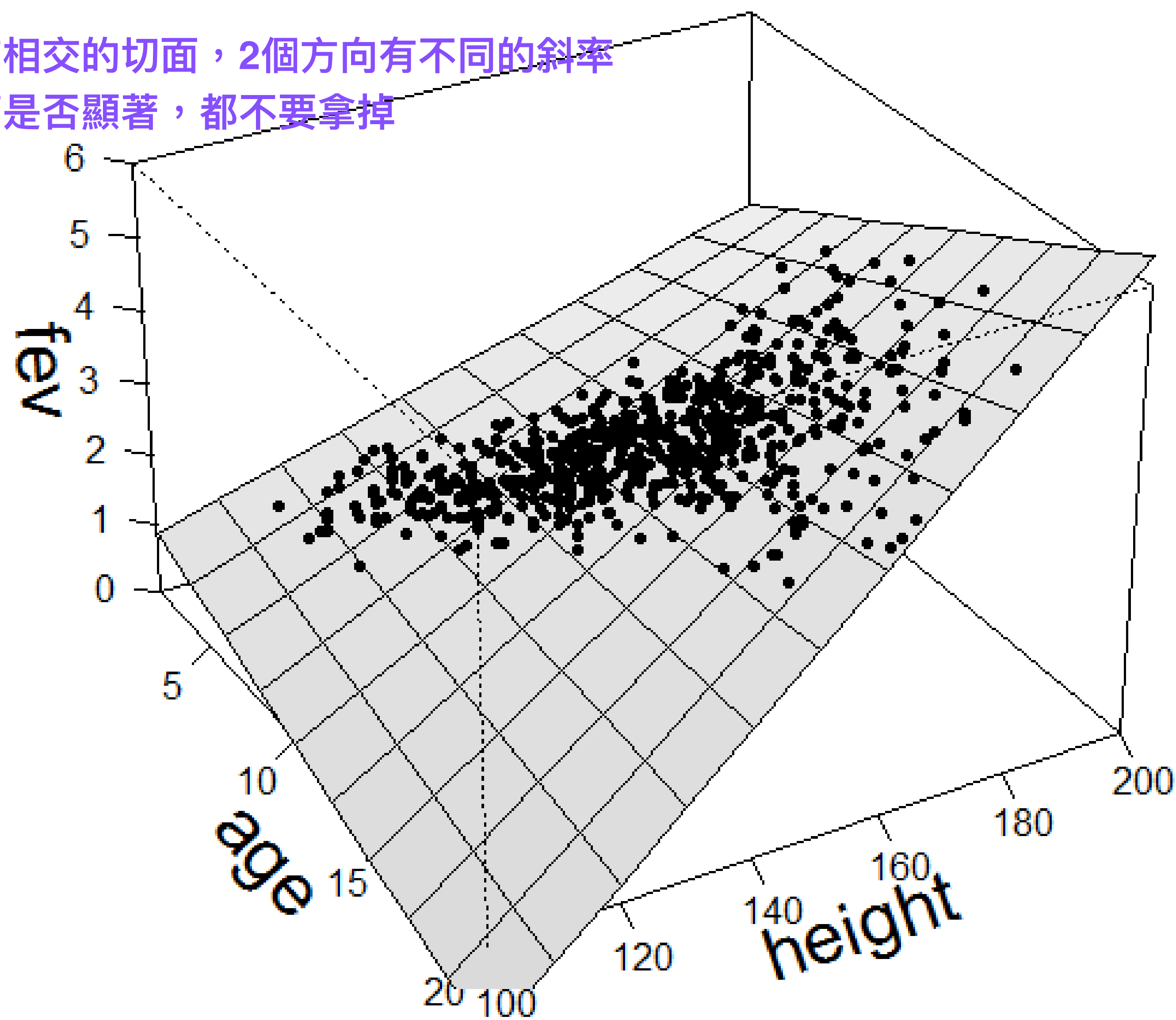


fev ~ age + height + age*height

放2個變向，年齡和身高 / FEV的關係，不見得都是正向的

任何一點有相交的切面，2個方向有不同的斜率

低系數不管是否顯著，都不要拿掉



Interaction Between Two Continuous Variables

```
> lm7<-lm(fev ~ age*height, data = FEV)
> summary(lm7)
```

曲度大要保留，曲度小當平面

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.7084499  0.5116033  -1.385    0.167
age         -0.4108097  0.0562053  -7.309 7.92e-13 ***
height       0.0182820  0.0034521   5.296 1.62e-07 ***
age:height   0.0029097  0.0003471   8.383 3.19e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 0.3997 on 650 degrees of freedom
Multiple R-squared:  0.7884,   Adjusted R-squared:  0.7875
F-statistic: 807.4 on 3 and 650 DF,  p-value: < 2.2e-16
```

將模型重新排列，年齡和FEv的關係不是固定的，需考量身高
對____歲的人而言，每增加一公分，其FEV會增加________

$$\widehat{fev} = -0.708 - 0.411age + 0.018height + 0.003age * height$$

- We usually only interpret the coefficient for the interaction term, as coefficients for age and height is a little tricky to interpret
- The equation can be re-arranged as:

$$\widehat{fev} = -0.708 + (-0.411 + 0.003height) * age + 0.018height$$

- This means that the effect of age on fev depends on height, i.e. for people with different body heights, the changes in their fev when they become 1 year older are *different*