(8) 統計模型與迴歸分析

吳漢銘
淡江大學 數學系
資料科學與數理統計組

http://www.hmwu.idv.tw

本章大綱

- R軟體裡的統計模型配適　。
- 簡單線性迴歸 (Simple Linear Regression).
- Extract Information from Model Objects.
- 簡單線性迴歸之模型檢測
- 模型選取

學習目標

- 熟悉R的統計模型，並了解其**`formula`**的寫法及意義。
- 能用R寫出簡單線性迴歸之參數估計與信賴區間之副程式。並能和**`lm`**之答案對照。
- 能畫出資料之二維散佈圖並加上迴歸線。能擷取**`lm`**的各項資訊。
- 能利用step做迴歸模型選取

統計模型配適 (Statistical Modeling)

四個問題:

1. Which of your variables is the **response variable** (反應變數)?
2. Which are the **explanatory variable** (解釋變數)?
3. Are the explanatory variables **continuous** (連續) or **categorical** (類別), or a **mixture** (混合) of both?
4. What kind of response variable do you have: **continuous** measurement, a **count**, a **proportion**, a **time** at death, or **category**?

配適統計模型的目的

- To determine the values of the **parameters** in a specific model that lead to the best fit of the model to the data.

解釋變數

The Explanatory Variable (x)

- **All x's are continuous**: Regression

例如:

$$\begin{aligned}
&\text{Simple linear regression: } y = \beta_0 + \beta_1 x + \epsilon\\
&\text{Multiple linear regression: } y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon\\
&\text{Polynomial regression: } y = \beta_0 + \beta_1 x + +\beta_2 x^2 + \cdots + \beta_d x^d + \epsilon\\
&\text{Nonlinear regression: } y = \theta_0 + \theta_1 (1 - e^{\theta_2 x}) + \epsilon
\end{aligned}$$

- **All x's are categorical**: Analysis of Variance (ANOVA, 變異數分析)

例如:

$$\begin{aligned}
y_{ij} &= \mu + \alpha_j + \epsilon_{ij}\\
\boldsymbol{y} &= \boldsymbol{A\theta} + \boldsymbol{\epsilon}
\end{aligned}$$

- **x's are both continuous and categorical:** Analysis of Covariance (ANCOVA)

例如:

$$y = \beta_0 + \beta_1 x + \theta z + \epsilon, \quad z = \{0, 1\}$$

反應變數 (1)

The Response Variable (y)

- **Continuous**: Normal Regression, ANOVA or ANCOVA
- **Binary**: Binary Logistic Analysis

例如:

$$P(y_i = 0) = 1 - \pi_i,\ P(y_i = 1) = \pi_i$$

$$\text{Logistic link function: } g(\pi) = \log(\frac{\pi}{1-\pi})$$

$$\text{Logistic regression: } \log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- **Ordinal**: proportional-odds model

例如:

$$\gamma_j(\mathbf{x}) = P(Y \leq j|\mathbf{x}), \quad \log\left(\frac{\gamma_j(\mathbf{x})}{1-\gamma_j(\mathbf{x})}\right) = \boldsymbol{\beta}^T\mathbf{x}$$

反應變數 (2)

風險函數

- **Count**: Log-Linear Models

例如: $Y \sim Poisson(\mu),\ \mu = E(Y), \quad \log \mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$

- **Time at death**: Survival Analysis

  - T: survival time with a density function $f(t)$.
  - $1 - F(t)$: survival function (i.e., $F(t) = \int_{-\infty}^{t} f(s)\, ds$).
  - $h(t) = \dfrac{f(t)}{1 - F(t)}$: hazard function.
  - $h(t)\delta t$: the probability of dying in the next small interval δt given survival to time t
  - Proportional-hazards model: $h(t; \mathbf{x}) = \lambda(t) \exp(\beta^T \mathbf{x})$

模式寫法 (Model Formulae in R)

- The structure of the model: `response.variable ~ explanatory.variables`
  - Example: `fm <- formula(y ~ x)`
  - Example: `lm(fm), lm(y ~ x);  aov(y ~ x); glm(y ~ x)`
- `~`: "is modelled as a function of"
  - Example: `lm(y ~ x)`
- `+` : inclusion of an explanatory variable in the model (not addition);
  - Example: `lm(y ~ x1 + x2)`
- `-` : deletion of an explanatory variable from the model (not subtraction);
  - Example: `lm(y ~ x1 - 1)`
- `*` : inclusion of explanatory variables and interactions (not multiplication);
  - Example: `lm(y ~ x1 * x2)`
- `/`: nesting of explanatory variables in the model (not division);
  - Example: `lm(y ~ x1 / x2)`

模式寫法 (Model Formulae in R)

- |: indicates conditioning (not 'or'), so that y ~ x | z is read as 'y as a function of x given z'.
  - Example: `lm(y ~ x1 | x2)`
- "`:`" : a colon denotes an interaction
  - `A:B` means the two-way interaction between `A` and `B`
  - `N:P:K:Mg` means the four-way interaction between `N, P, K` and `Mg`.
- `A*B*C` is the same as `A+B+C+A:B+A:C+B:C+A:B:C`
- `A/B/C` is the same as `A+B%in%A+C%in%B%in%A`
- `(A+B+C)^3` is the same as `A*B*C`
- `(A+B+C)^2` is the same as `A*B*C – A:B:C`

```
> ##Create a formula for a model with a large number of variables:
> xnam <- paste("x", 1:25, sep="")
> (fmla <- as.formula(paste("y ~ ", paste(xnam, collapse= "+"))))
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 +
    x12 + x13 + x14 + x15 + x16 + x17 + x18 + x19 + x20 + x21 +
    x22 + x23 + x24 + x25
```

Model Formula 例子1

Table 9.3. Examples of R model formulae. In a model formula, the function I case i) stands for 'as is' and is used for generating sequences I(1:10) or calculating quadratic terms I(x^2).

Model	Model formula	Comments
Null	y ~ 1	1 is the intercept in regression models, but here it is the overall mean *y*
Regression	y ~ x	*x* is a continuous explanatory variable
Regression through origin	y ~ x-1	Do not fit an intercept **y ~ 0 + x**
One-way ANOVA	y ~ sex	sex is a two-level categorical variable
One-way ANOVA	y ~ sex-1	as above, but do not fit an intercept (gives two means rather than a mean and a difference)
Two-way ANOVA	y ~ sex + genotype	genotype is a four-level categorical variable
Factorial ANOVA	y ~ N * P * K	*N*, *P* and *K* are two-level factors to be fitted along with all their interactions

Source: Crawley, M. J. , 2007, The R Book, Wiley.

Model Formula 例子2

Table 9.3. (Continued)

Model	**Model formula**	**Comments**
Three-way ANOVA	y ~ N*P*K – N:P:K	As above, but don't fit the three-way interaction
Analysis of covariance	y ~ x + sex	A common slope for *y* against *x* but with two intercepts, one for each sex
Analysis of covariance	y ~ x * sex	Two slopes and two intercepts
Nested ANOVA	y ~ a/b/c	Factor *c* nested within factor *b* within factor *a*
Split-plot ANOVA	y ~ a*b*c+Error(a/b/c)	A factorial experiment but with three plot sizes and three different error variances, one for each plot size
Multiple regression	y ~ x + z	Two continuous explanatory variables, flat surface fit
Multiple regression	y ~ x * z	Fit an interaction term as well (x + z + x:z)

Source: Crawley, M. J. , 2007, The R Book, Wiley.

Model Formula 例子 3

Table 9.3. (Continued)

Model	**Model formula**	**Comments**
Multiple regression	y ~ x + I(x^2) + z + I(z^2)	Fit a quadratic term for both x and z
Multiple regression	y <- poly(x,2) + z	Fit a quadratic polynomial for x and linear z
Multiple regression	y ~ (x + z + w)^2	Fit three variables plus all their interactions up to two-way
Non-parametric model	y ~ s(x) +s(z)	y is a function of smoothed x and z in a generalized additive model
Transformed response and explanatory variables	log(y) ~ I(1/x) + sqrt(z)	All three variables are transformed in the model

the function I case i) stands
for 'as is' and is used for generating sequences I(1:10)
or calculating quadratic terms I(x^2).

Source: Crawley, M. J. , 2007, The R Book, Wiley.

Statistical Models in R

lm fits a linear model with normal errors and constant variance; generally this is used for regression analysis using continuous explanatory variables.

aov fits analysis of variance with normal errors, constant variance and the identity link; generally used for categorical explanatory variables or ANCOVA with a mix of categorical and continuous explanatory variables.

glm fits generalized linear models to data using categorical or continuous explanatory variables, by specifying one of a family of **error structures** (e.g. Poisson for count data or binomial for proportion data) and a particular **link function**.

gam fits generalized additive models

lme and lmer fit linear mixed-effects models

nls fits a non-linear regression model via least squares

nlme fits a specified non-linear function in a mixed-effects model

loess fits a local regression model

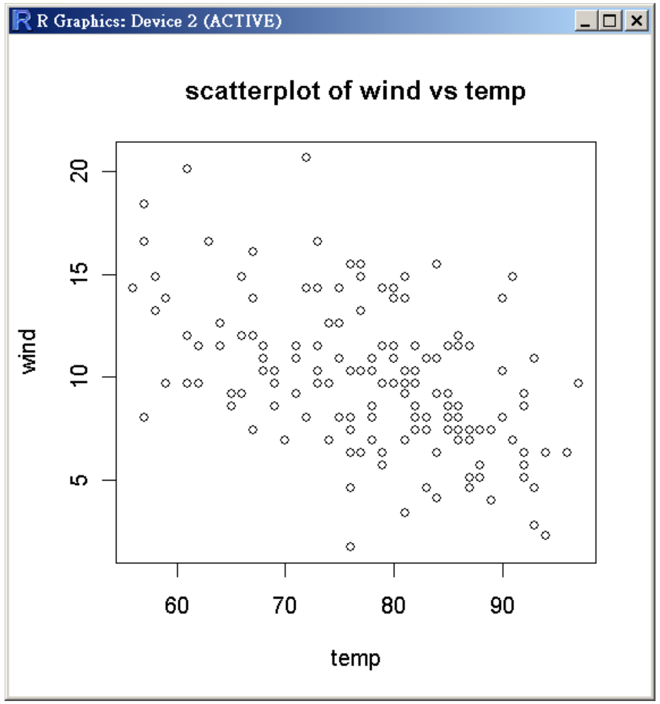
tree fits a regression tree model using binary recursive partitioning

Crawley, M. J. , 2007, The R Book, Wiley.

簡單線性迴歸 (Simple Linear Regression)

$$y = \beta_0 + \beta_1 x + \epsilon$$
$$E(\epsilon) = 0$$
$$Var(\epsilon) = \sigma^2$$

$$E(y|x) = \beta_0 + \beta_1 x$$
$$Var(y|x) = Var(\beta_0 + \beta_1 x + \epsilon) = \sigma^2$$




```
> wind <- airquality$Wind
> temp <- airquality$Temp
> plot(temp, wind, main="scatterplot of wind vs temp")
```

- beta_0 (intercept), beta_1 (slope): parameters to be estimated from observed data.
- Random errors (epsilon): mean zero and unknown variance (sigma^2).
- The variance in y is constant (i.e. the variance does not change as y gets bigger).

參數估計: 最小平方法

$$(y_1, x_1), \cdots, (y_n, x_n)$$

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

```
> y <- airquality$Wind
> x <- airquality$Temp
> xbar <- mean(x) ; xbar
[1] 77.88235
> ybar <- mean(y) ; ybar
[1] 9.957516

> beta1.num <- sum((x-xbar)*(y-ybar))
> beta1.den <- sum((x-xbar)^2)
> (beta1.hat <- beta1.num/beta1.den)
[1] -0.1704644

> (beta0.hat <- ybar-beta1.hat*xbar)
[1] 23.23369
> yhat <- beta0.hat + beta1.hat * x
```

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

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

$$S_{xy} = \sum_{i=1}^{n} y_i (x_i - \bar{x})$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

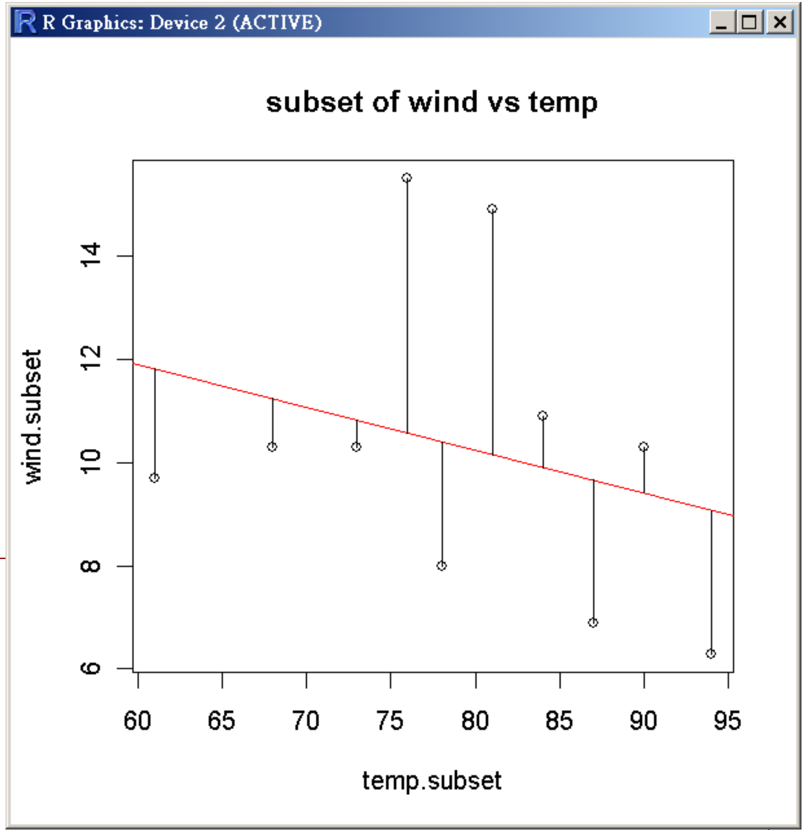
```
> Sxy <- sum(y*(x-xbar)) ; Sxy
[1] -2321.365
> Sxx <- sum((x-xbar)^2) ; Sxx
[1] 13617.88
> Syy <- sum((y-ybar)^2) ; Syy
[1] 1886.554
> beta1.hat2 <- Sxy/Sxx ; beta1.hat2
[1] -0.1704644
```

最小平方法

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

$$e_i = y_i - \hat{y}_i$$

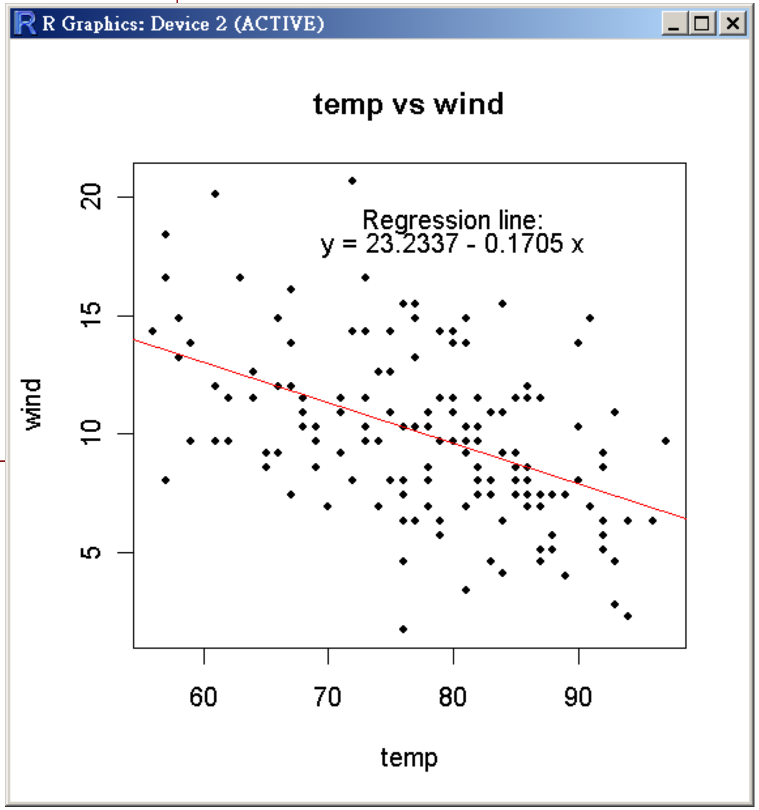
和 **summary(lm(y~x))** 比較




```
> wind <- airquality$Wind
> temp <- airquality$Temp

> n <- length(wind)
> index <- sample(1:n, 10)
> wind.subset <- wind[index]
> temp.subset <- temp[index]

> plot(wind.subset~temp.subset, main="subset of wind vs temp")
> subset.lm <- lm(wind.subset~temp.subset)
> abline(subset.lm, col="red")
> segments(temp.subset, fitted(subset.lm), temp.subset, wind.subset)
```

Find the Least Squares Fit

```
> model.fit <- lsfit(temp, wind)
> ls.print(model.fit)
Residual Standard Error=3.1422
R-Square=0.2098
F-statistic (df=1, 151)=40.0795
p-value=0

          Estimate Std.Err t-value Pr(>|t|)
Intercept  23.2337  2.1124 10.9987        0
X          -0.1705  0.0269 -6.3308        0


> plot(temp, wind, main="temp vs wind", pch=20)
> abline(model.fit, col="red")
> text(80,19, "Regression line:")
> text(80,18, "y = 23.2337 - 0.1705 x")
```

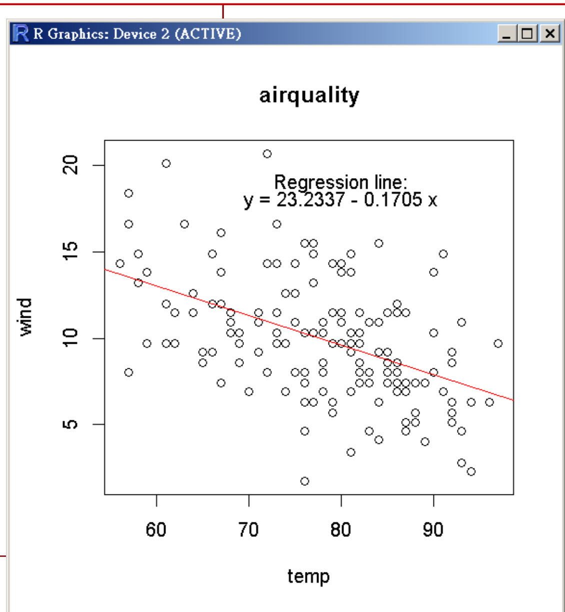


Fit A Linear Model: lm

```
> my.model <- lm(wind ~ temp)
> my.model

Call:
lm(formula = wind ~ temp)

Coefficients:
(Intercept)         temp
    23.2337      -0.1705

> summary(my.model)

Call:
lm(formula = wind ~ temp)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5784 -2.4489 -0.2261  1.9853  9.7398

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.23369    2.11239  10.999  < 2e-16 ***
temp        -0.17046    0.02693  -6.331 2.64e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.142 on 151 degrees of freedom
Multiple R-squared: 0.2098,     Adjusted R-squared: 0.2045
F-statistic: 40.08 on 1 and 151 DF,  p-value: 2.642e-09
```

```
> my.aov <- aov(my.model)
> summary(my.aov)
             Df  Sum Sq Mean Sq F value    Pr(>F)
temp          1  395.71  395.71  40.080 2.642e-09 ***
Residuals   151 1490.84    9.87
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.
```

```
plot(wind ~ temp, main="airquality")
abline(my.model, col="red")
text(80,19, "Regression line:")
text(80,18, "y = 23.2337 - 0.1705 x")
```




```
> plot(my.model, which=1:6)
```

Sum of Squares and ANOVA Table

$$e_i = y_i - \hat{y}_i$$

$$SS_E = \sum_{i=1}^{n} e_i^2 \qquad MS_E = \frac{SS_E}{n-2} = \hat{\sigma^2}$$

$$SS_R = \hat{\beta}_1 S_{xy} \qquad MS_R = SS_R/1$$

$$F_0 = MS_R/MS_E$$

```
> n <- length(wind)
> e <- y-yhat
> SSE <- sum(e^2) ; SSE
[1] 1490.844
> MSE <- SSE/(n-2) ; MSE
[1] 9.873137
> SSR <- beta1.hat*Sxy ; SSR
[1] 395.7101
> MSR <- SSR/1 ; MSR
[1] 395.7101
> SST <- SSR + SSE ; SST
[1] 1886.554
> Syy
[1] 1886.554
> F0 <- MSR/MSE; F0
[1] 40.07947
```

The ANOVA Table for Regression

Source	SS *(Sum of Squares, the numerator of the variance)*	DF *(the denominator)*	MS *(Mean Square, the variance)*	F
Regression *(or Model)*	$SSR = \sum_{i=1}^{n} ((\hat{\beta}_0 + \hat{\beta}_1 x_i) - \bar{y})^2$	*2-1=1*	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Error	$SSE = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$	*n-2*	$MSE = \frac{SSE}{n-2}$	
Total	$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$	*n-1*		

課堂練習1: 估計量

- 用R寫出以下估計量，並與上述例子的答案比較。

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{1}{n}(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{2} x_i)^2}$$

$$SS_E = SS_T - SS_R$$

$$SS_E = S_{yy} - \hat{\beta}_1 S_{xy}$$

$$R^2 = \frac{SS_R}{S_{yy}} = 1 - \frac{SS_E}{S_{yy}}$$

決定係數 Coefficient of Determination

信賴區間

$100(1-\alpha)\%$ confident interval on the intercept β_0.

$$E(\hat{\beta}_0) = \beta_0 \qquad se(\hat{\beta}_0) = \sqrt{MS_E(1/n + \bar{x}^2/S_{xx})}$$

$$\hat{\beta}_0 - t_{\alpha/2,n-1} se(\hat{\beta}_0) \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-1} se(\hat{\beta}_0)$$

$100(1-\alpha)\%$ confident interval on the slope β_1.

$$E(\hat{\beta}_1) = \beta_1 \qquad se(\hat{\beta}_1) = \sqrt{MS_E/S_{xx}}$$

$$\hat{\beta}_1 - t_{\alpha/2,n-1} se(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-1} se(\hat{\beta}_1)$$

```
> alpha <- 0.05
> se.beta0 <- sqrt(MSE*(1/n+xbar^2/Sxx)) ; se.beta0
[1] 2.112395
> tstar <- qt(alpha/2, n-1)* se.beta0
> CI.beta0 <- beta0.hat + c(-tstar*se.beta0, tstar*se.beta0) ; CI.beta0
[1] 32.04965 14.41772
```

```
> se.beta1 <- sqrt(MSE/Sxx) ; se.beta1
[1] 0.02692606
> tstar <- qt(alpha/2, n-1)* se.beta1
> CI.beta1 <- beta1.hat + c(-tstar*se.beta0, tstar*se.beta1); CI.beta1
[1] -0.0580900 -0.1718968
```

課堂練習2: 信賴區間

- 用R寫出以下估計量，並用以上的例子算出答案。

$100(1-\alpha)\%$ confident interval on σ^2.

$$\frac{(n-2)MS_E}{\chi^2_{\alpha/2,n-2}} \le \sigma^2 \le \frac{(n-2)MS_E}{\chi^2_{1-\alpha/2,n-2}}$$

$100(1-\alpha)\%$ confident interval on

the mean response at the point $x = x_0$.

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{MS_E\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)} \le E(y|x_0) \le \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{MS_E\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

Generic Functions

```
> my.model <- lm(wind ~ temp)
> summary(my.model)
```

- **`summary`**: produces parameter estimates and standard errors from **`lm`**, and ANOVA tables from **`aov`**.
- **`plot`**: produces diagnostic plots for model checking, including residuals against fitted values, influence tests, etc.
- **`update`**: is used to modify the last model fit; it saves both typing effort and computing time.
- **`predict`**: uses information from the fitted model to produce smooth functions for plotting a line through the scatterplot of your data.
- **`fitted`**: gives the fitted values, predicted by the model for the values of the explanatory variables included.
- **`resid`**: gives the residuals.

Extract Information from Model Objects

方法一: by names

```
> coef(my.model)
(Intercept)        temp
 23.2336881  -0.1704644

> vcov(my.model)
             (Intercept)          temp
(Intercept)   4.46221130 -0.0564656925
temp         -0.05646569  0.0007250127
```

```
> my.model <- lm(wind ~ temp)
> summary(my.model)

Call:
lm(formula = wind ~ temp)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5784 -2.4489 -0.2261  1.9853  9.7398

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.23369    2.11239  10.999  < 2e-16 ***
temp        -0.17046    0.02693  -6.331 2.64e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.142 on 151 degrees of freedom
Multiple R-squared: 0.2098,     Adjusted R-squared: 0.2045
F-statistic: 40.08 on 1 and 151 DF,  p-value: 2.642e-09
```

Extract Information from Model Objects

方法二: with list subscripts

```
> summary(my.model)[[1]]  #my.model formula
lm(formula = wind ~ temp)

> summary(my.model)[[2]]  #attributes of the objects
wind ~ temp
attr(,"variables")
list(wind, temp)
attr(,"factors")
     temp
wind    0
temp    1
attr(,"term.labels")
[1] "temp"
attr(,"order")
[1] 1
attr(,"intercept")
[1] 1
attr(,"response")
[1] 1
attr(,".Environment")
<environment: R_GlobalEnv>
attr(,"predvars")
list(wind, temp)
attr(,"dataClasses")
     wind      temp
"numeric" "numeric"
```

```
> length(summary(my.model))
[1] 11
> names(summary(my.model))
 [1] "call"          "terms"         "residuals"     "coefficients"
 [5] "aliased"       "sigma"         "df"            "r.squared"
 [9] "adj.r.squared" "fstatistic"    "cov.unscaled"

> summary(my.model)$sigma
[1] 3.142155
> summary(my.model)[[6]]
[1] 3.142155

> length(summary(my.model)[[1]])
[1] 2
> length(summary(my.model)[[2]])
[1] 3
> length(summary(my.model)[[3]])
[1] 153
```

Extract Information from Model Objects

方法二: with list subscripts

```
> summary(my.model)[[3]]  #residuals for data points
          1           2           3           4           5           6
-4.41257055 -2.96024835  1.98068054 -1.16489276  0.61232059  2.91696501
...
145         146         147         148         149         150
-1.93071279  0.87393162 -1.17164167  4.10557168 -4.40117723  3.09207386
        151         152         153
 3.85114498 -2.27839058 -0.14210611

> summary(my.model)[[4]]  #parameters table
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 23.2336881 2.11239468 10.998744 4.901351e-21
temp        -0.1704644 0.02692606 -6.330835 2.641597e-09

> summary(my.model)[[4]][[1]]  #intercept
[1] 23.23369

> summary(my.model)[[4]][[2]]  #slope,.... summary(my.model)[[4]][[28]]
[1] -0.1704644
```

```
> str(summary(my.model)[[4]])
 num [1:2, 1:4] 23.2337 -0.1705  2.1124  0.0269 10.9987 ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:2] "(Intercept)" "temp"
  ..$ : chr [1:4] "Estimate" "Std. Error" "t value" "Pr(>|t|)"
```

Extract Information from Model Objects

方法二: with list subscripts

```
> summary(my.model)[[5]]  #whether the fit should be returned.
(Intercept)        temp
      FALSE       FALSE

> summary(my.model)[[6]]  #residual standard error
[1] 3.142155

> summary(my.model)[[7]]  #the number of rows in the summary.lm table.
[1]   2 151   2

> summary(my.model)[[8]]  #r square, the fraction of the total variation in the response
    variable that is explained by the my.model.
[1] 0.2097529

> summary(my.model)[[9]]  #adjusted r square
[1] 0.2045195

> summary(my.model)[[10]]  #F ratio information
    value     numdf     dendf
 40.07947   1.00000 151.00000

> summary(my.model)[[11]]  #correlation matrix of the parameter estimates.
             (Intercept)          temp
(Intercept)  0.451954754 -5.719124e-03
temp        -0.005719124  7.343286e-05
```

Extract Information from Model Objects

方法三: using $

```
> my.model <- lm(wind ~ temp)
> names(my.model)
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"

> model$coefficients
> model$fitted.values
> model$residuals
```

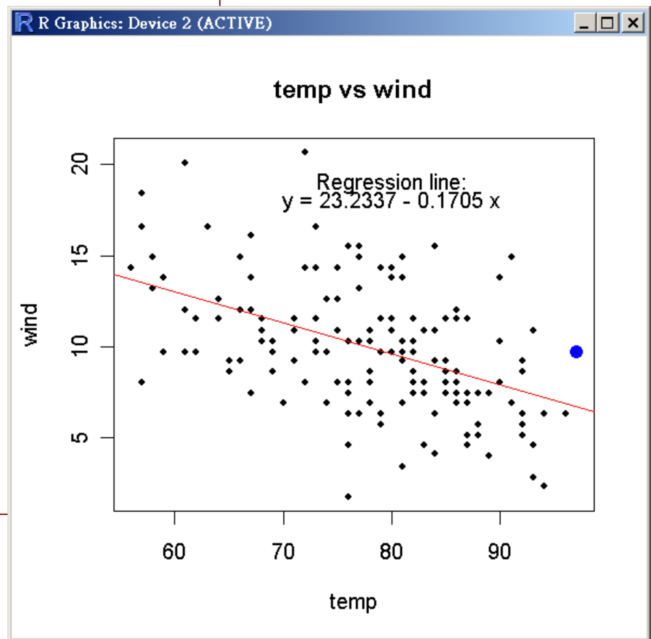
依此類推...

```
> summary.aov(my.model)
> summary.aov(my.model)[[1]][[1]]~
> summary.aov(my.model)[[1]][[5]]
```

使用子集合 (Using Subset)

- Investigate how much a influence point affected the parameter estimates and their standard error.
- Repeat the statistical modeling but leave out the point in question, using subset.

```
> new.model <- update(my.model, subset=(temp!=max(temp)))
> summary(new.model)

Call:
lm(formula = wind ~ temp, subset = (temp != max(temp)))

Residuals:
    Min      1Q  Median      3Q     Max 
-8.5663 -2.3871 -0.2027  1.9662  9.7344 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  23.5529     2.1382  11.015  < 2e-16 ***
temp         -0.1748     0.0273  -6.403 1.85e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 3.143 on 150 degrees of freedom
Multiple R-squared: 0.2147,     Adjusted R-squared: 0.2094 
F-statistic:    41 on 1 and 150 DF,  p-value: 1.847e-09
```



課堂練習：

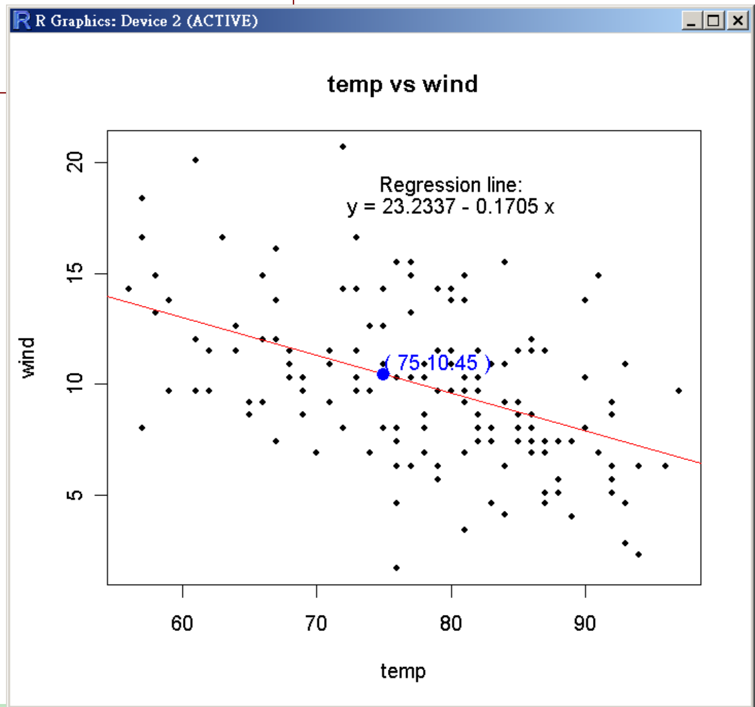
- 將要刪除的點在二維散佈圖上標出來。
- 更新二維散佈圖及Regression Fit。

預測 (Prediction)

```
> summary(wind)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.700   7.400   9.700   9.958  11.500  20.700 
> summary(temp)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  56.00   72.00   79.00   77.88   85.00   97.00 

> predict(my.model, list(temp=75))
[1] 10.44886
> predict(my.model, list(temp=c(66, 80,100)))
        1         2         3 
11.983035  9.596533  6.187244 
```

課堂練習:

- 將predict出來的值在二維散佈圖上標出來。



統計模型檢測 (Model Checking in R)

- After fitting a model to data we need to investigate how well the model describes the data.
- In particular, we should look to see if there are any systematic trends in the goodness of fit.
- Fit a linear regression (`lm`) to these data and then use model-checking plots (`plot`) to investigate the adequacy of that model.

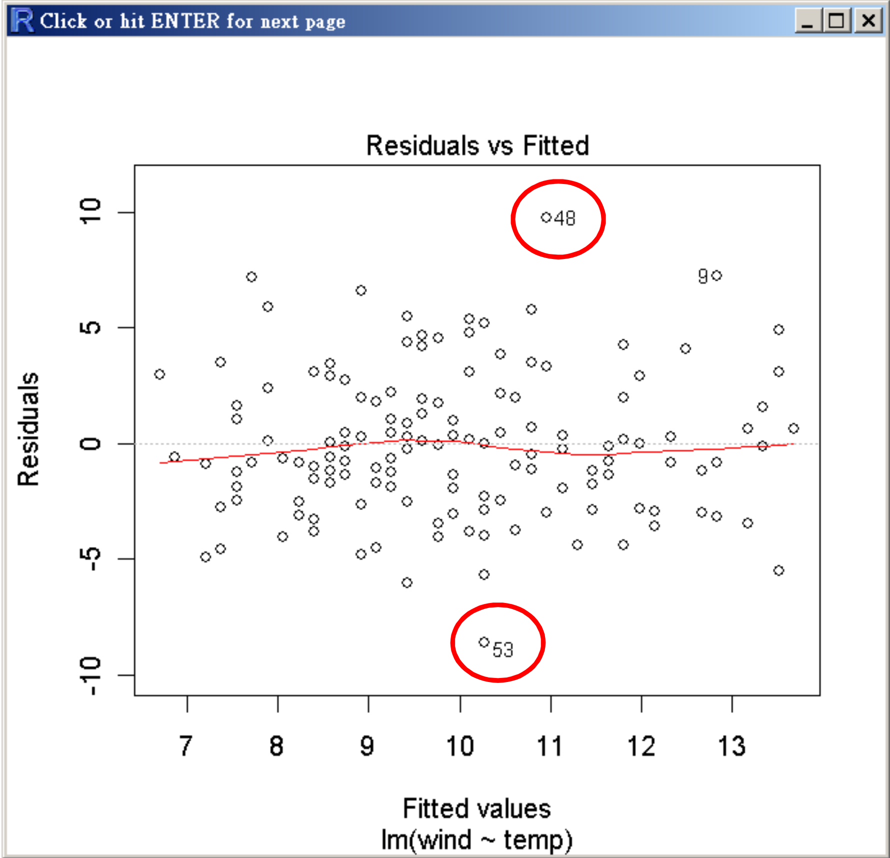
```
> ?plot.lm
```

```
> wind <- airquality$Wind
> temp <- airquality$Temp
> my.model <- lm(wind ~ temp)
> plot(my.model, which=1:6)
Waiting to confirm page change...
Waiting to confirm page change...
```

1. 殘差vs. 估計值 (Residuals vs Fitted Values)

default

This plot should be
with no pattern of any sort.

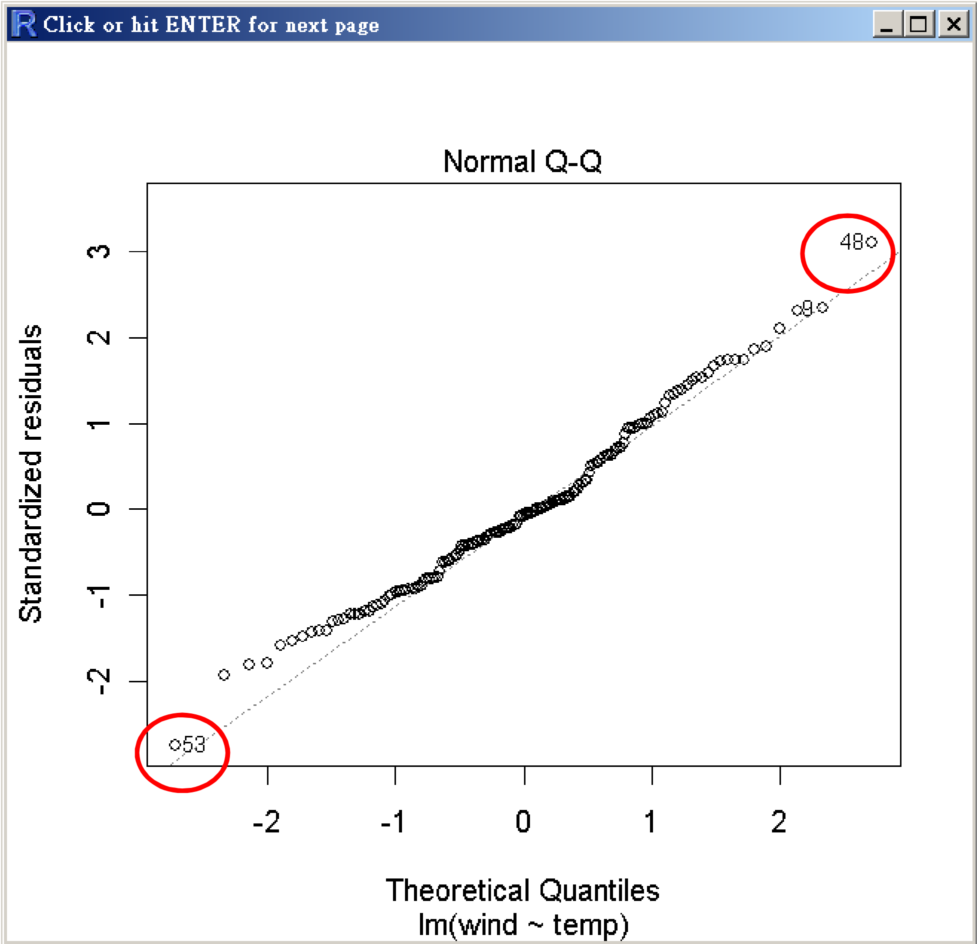



```
> plot(fitted(my.model), residuals(my.model),xlab="Fitted values",
ylab="Residuals")
> abline(h=0, lty=2)
```

課堂練習: 將Residuals大於±6的點標出來(顏色為紅色)。

2. 常態QQ圖 (Normal QQ-plot)

default

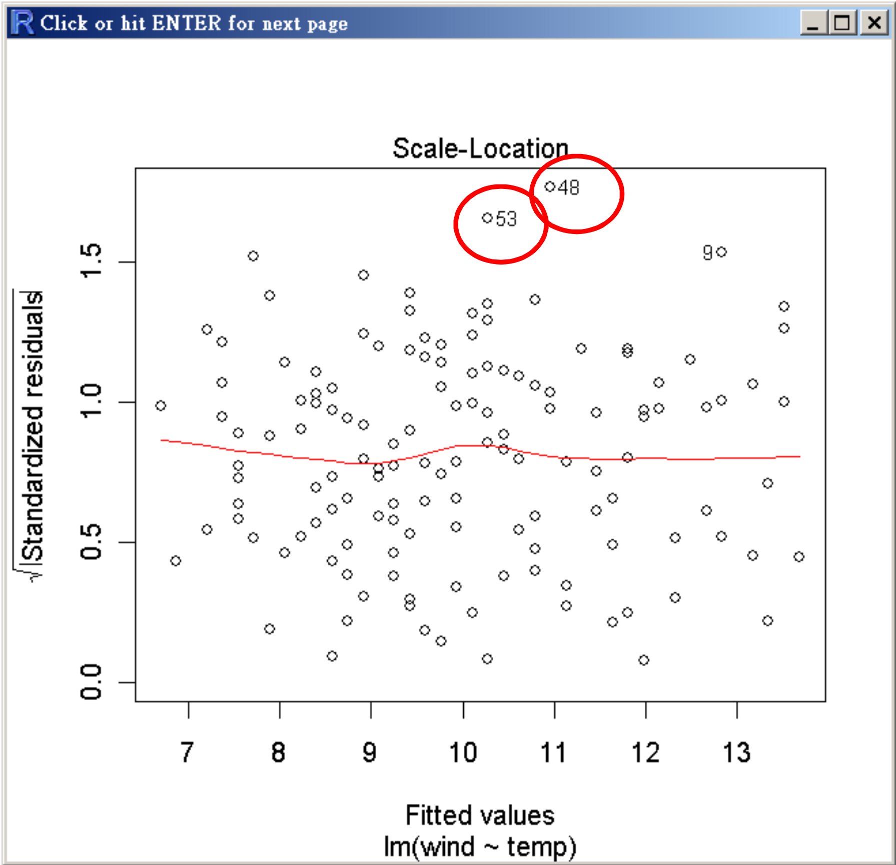



```
> qqnorm(residuals(my.model))
> qqline(residuals(my.model))
```

3. 尺度-位置圖 (A Scale-Location Plot)

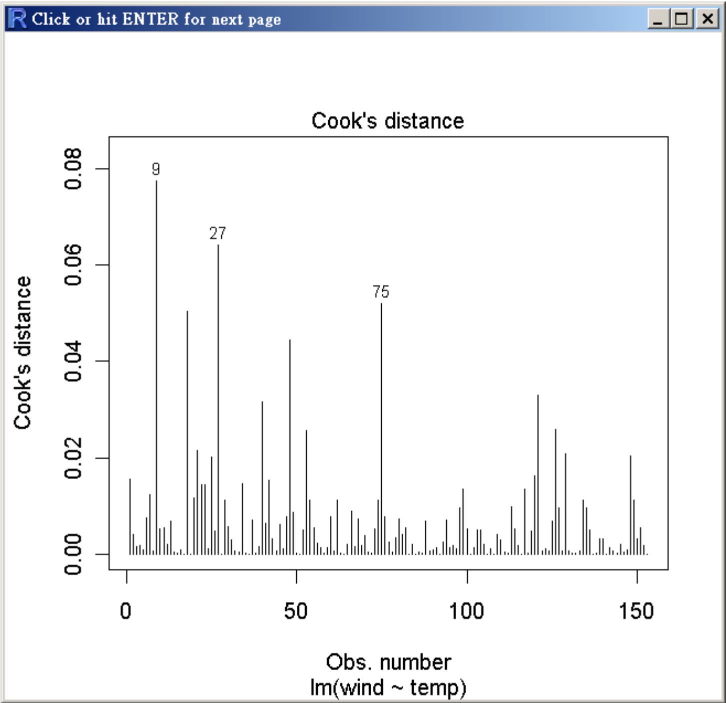
- A scale-loaction plot of sqrt(abs(residuals)) against fitted values.
- This is like a positive-valued version of the first graph; it is good for detecting non-constancy of variance (heteroscedasticity).

default



4. Plot of Cook's Distance vs Row Labels

- Cook's distance measures the effect of deleting a given observation.
- Cook's distance is a measure of the squared distance between the least square estimate based on all n points β and the estimate obtained by deleting the *i*th points $\beta_{(i)}$.
- Points with a Cook's distance of 1 or more are considered to be influential.



$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})}{pMS_E}$$

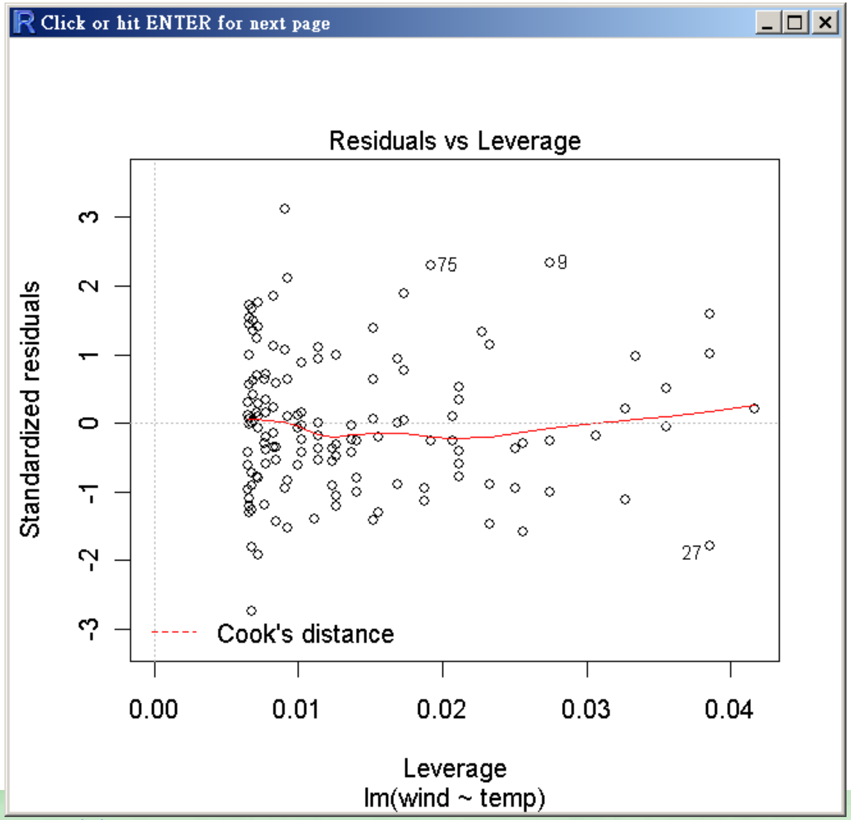
課堂練習:

- 算出Cook's Distance。
- 畫出Cook's Distance vs. Row Labels的散佈圖。
- 標出前三大Cook's Distance值所在位置。

5. Plot of Residuals vs Leverages

- Outliers in the response variable are called outliers.
- Outliers with respect to the predictors are called leverage points.
- For the regression, it is the points that have large leverage are important.
- Points that have small leverage "do not count" in the regression – we could move them or remove them from the data and the regression line does not change very much.

default



$$\mathrm{Le}_i = \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

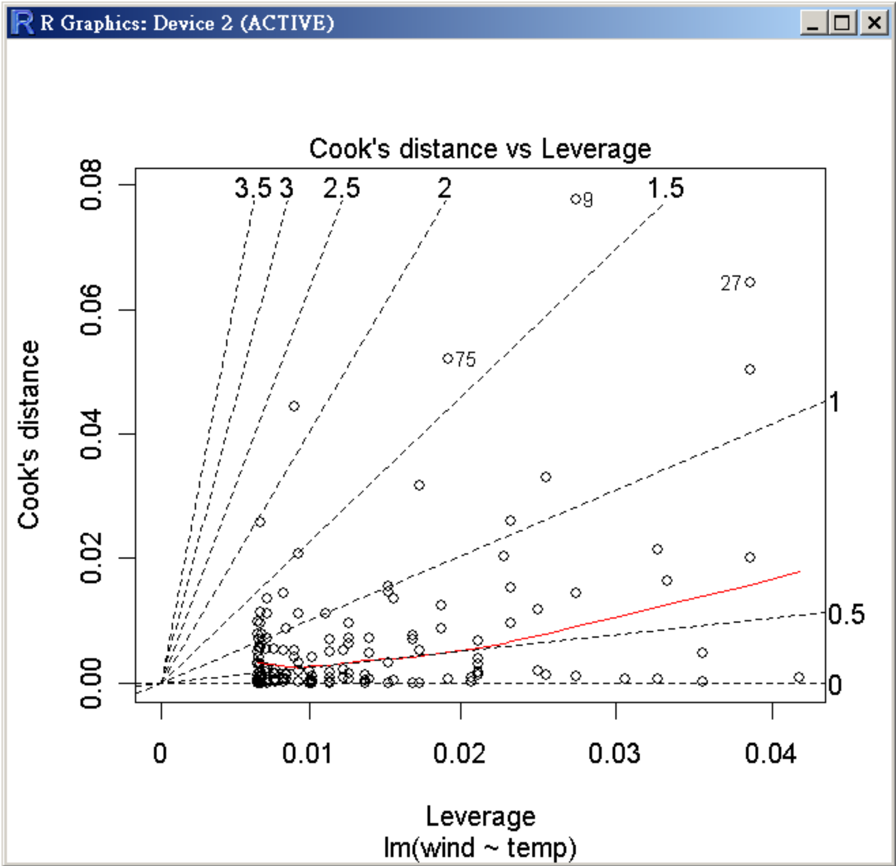
$$\hat{\beta}_1 = \sum_{i=1}^{n} \mathrm{Le}_i \frac{(y_i - \bar{y})}{(x_i - \bar{x})}$$

課堂練習 2:

- 算出Leverages 。
- 將Residuals標準化。
- 畫出Residuals標準化 vs. Leverages 的散佈圖。
- 標出前三大Leverages值所在位置。

6. Cook's Distance vs Leverage

- In the Cook's distance vs leverage/(1-leverage) plot, contours of standardized residuals that are equal in magnitude are lines through the origin.



模型選取/變數選取

Swiss Fertility and Socioeconomic Indicators (1888) Data

```
> head(swiss)
             Fertility Agriculture Examination Education Catholic Infant.Mortality
Courtelary        80.2        17.0          15        12     9.96             22.2
Delemont          83.1        45.1           6         9    84.84             22.2
Franches-Mnt      92.5        39.7           5         5    93.40             20.2
Moutier           85.8        36.5          12         7    33.77             20.3
Neuveville        76.9        43.5          17        15     5.16             20.6
Porrentruy        76.1        35.3           9         7    90.57             26.6
```

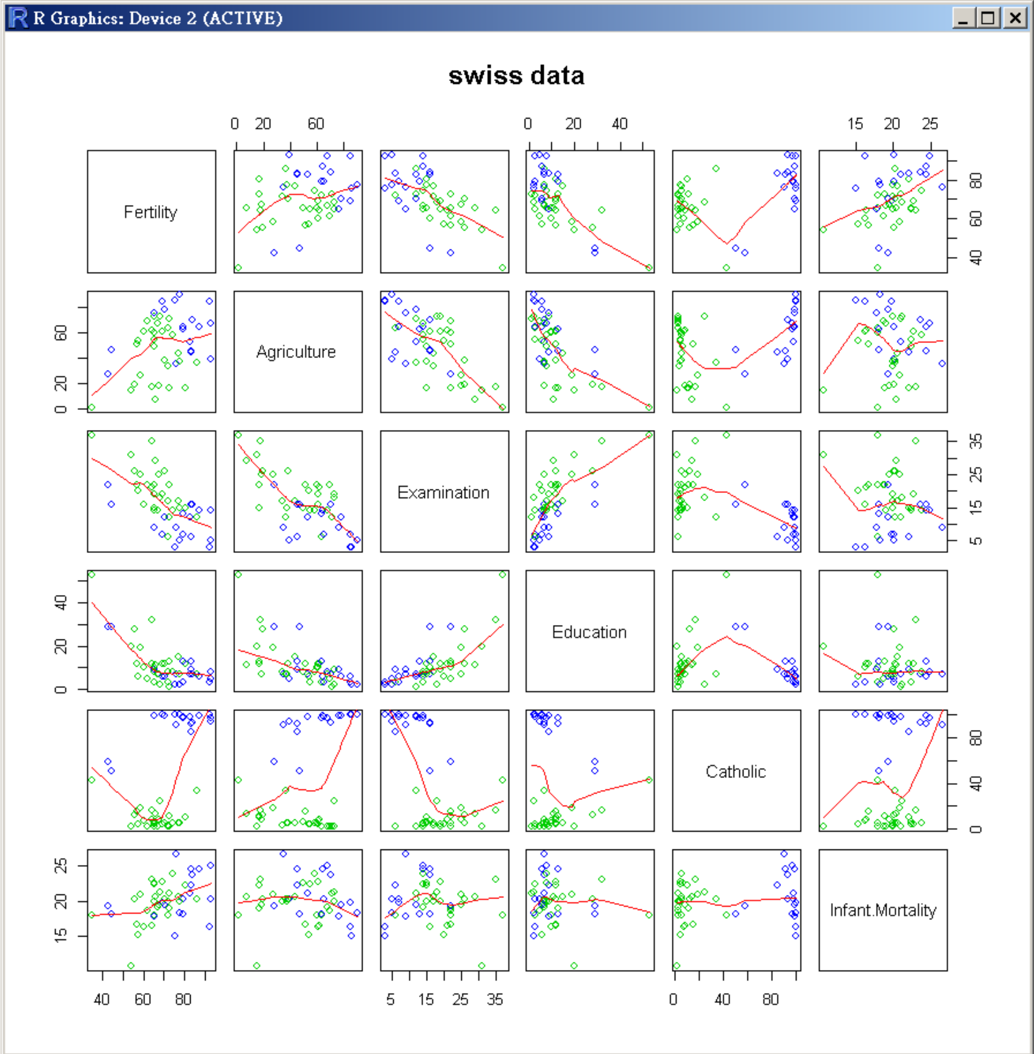
A data frame with 47 observations on 6 variables, each of which is in percent, i.e., in [0, 100].

[,1]	Fertility	Ig, ‘common standardized fertility measure’
[,2]	Agriculture	% of males involved in agriculture as occupation
[,3]	Examination	% draftees receiving highest mark on army examination
[,4]	Education	% education beyond primary school for draftees.
[,5]	Catholic	% ‘catholic’ (as opposed to ‘protestant’).
[,6]	Infant.Mortality	live births who live less than 1 year.

All variables but ‘Fertility’ give proportions of the population.

散佈圖矩陣

```
> pairs(swiss, panel = panel.smooth, main = "swiss data",
+         col = 3 + (swiss$Catholic > 50))
```



配適多重迴歸模型: lm

.=ALL

```
> summary(my.lm <- lm(Fertility ~ ., data = swiss))

Call:
lm(formula = Fertility ~ ., data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.2743  -5.2617   0.5032   4.1198  15.3213 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
Agriculture      -0.17211    0.07030  -2.448  0.01873 *  
Examination      -0.25801    0.25388  -1.016  0.31546    
Education        -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic          0.10412    0.03526   2.953  0.00519 ** 
Infant.Mortality  1.07705    0.38172   2.822  0.00734 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared:  0.7067,    Adjusted R-squared:  0.671 
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

逐步迴歸變數篩選: step

AIC (Akaike information criterion)常用來作為模型選取的準則。其值越小，代表模型的解釋能力越好(用的變數越少，或是誤差平方和越小)。

$$AIC = \ln\left(\frac{ESS}{n}\right) + \frac{2p}{n}, \quad ESS = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

語法:

```
step(object, scope, scale = 0, direction
= c("both", "backward", "forward"), trace
= 1, keep = NULL, steps = 1000, k = 2,
...)
```

```
> smy.lm <- step(my.lm)
Start:  AIC=190.69
Fertility ~ Agriculture + Examination + Education + Catholic +
    Infant.Mortality

                   Df Sum of Sq    RSS    AIC
- Examination       1     53.03 2158.1 189.86
<none>                          2105.0 190.69
- Agriculture       1    307.72 2412.8 195.10
- Infant.Mortality  1    408.75 2513.8 197.03
- Catholic          1    447.71 2552.8 197.75
- Education         1   1162.56 3267.6 209.36

Step:  AIC=189.86
Fertility ~ Agriculture + Education + Catholic + Infant.Mortality

                   Df Sum of Sq    RSS    AIC
<none>                          2158.1 189.86
- Agriculture       1    264.18 2422.2 193.29
- Infant.Mortality  1    409.81 2567.9 196.03
- Catholic          1    956.57 3114.6 205.10
- Education         1   2249.97 4408.0 221.43
```

最後選取的模型

```
> summary(smy.lm)

Call:
lm(formula = Fertility ~ Agriculture + Education + Catholic +
    Infant.Mortality, data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max
-14.6765  -6.0522   0.7514   3.1664  16.1422

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      62.10131    9.60489   6.466 8.49e-08 ***
Agriculture      -0.15462    0.06819  -2.267  0.02857 *
Education        -0.98026    0.14814  -6.617 5.14e-08 ***
Catholic          0.12467    0.02889   4.315 9.50e-05 ***
Infant.Mortality  1.07844    0.38187   2.824  0.00722 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.168 on 42 degrees of freedom
Multiple R-squared:  0.6993,    Adjusted R-squared:  0.6707
F-statistic: 24.42 on 4 and 42 DF,  p-value: 1.717e-10
```