應用生物統計學
Applied Biostatistics
線性迴歸
Linear Regression

林菀俞 (Wan-Yu Lin)
台大公衛學院流行病學與預防醫學研究所

2012.10.23

為什麼要學 linear regression？

- 想知道 rs35682 上 ‘A’ allele 的個數(0, 1, or 2)和身體質量指數(body-mass index)之間的關係為何？ => simple linear regression
- 想知道 rs35682 上 ‘A’ allele 的個數(0, 1, or 2)和身體質量指數(body-mass index)之間的關係為何，但年紀、性別可能影響此二者間的關係 => multiple linear regression
- Functional relation vs. Statistical relation
- BMI data set (homework)

線性迴歸模式

$$\boxed{Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i}$$

- Y_i ：依變項(dependent variable, response variable)
- $X_{i1}, X_{i2}, \cdots, X_{i,p-1}$ ：自變項(independent variables, explanatory variables, predictors, covariates)
- ε_i ：隨機誤差(random error)，假設 $N\left(0, \sigma^2\right)$
- 假設 ε_i 與 ε_j 之間無相關
- $\beta_0, \beta_1, \beta_2, \cdots, \beta_{p-1}$：迴歸係數(regression coefficients)，未知需估計

$$E\left(Y_i\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1}$$

線性迴歸模式的矩陣表示法

$$\boxed{Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i}$$

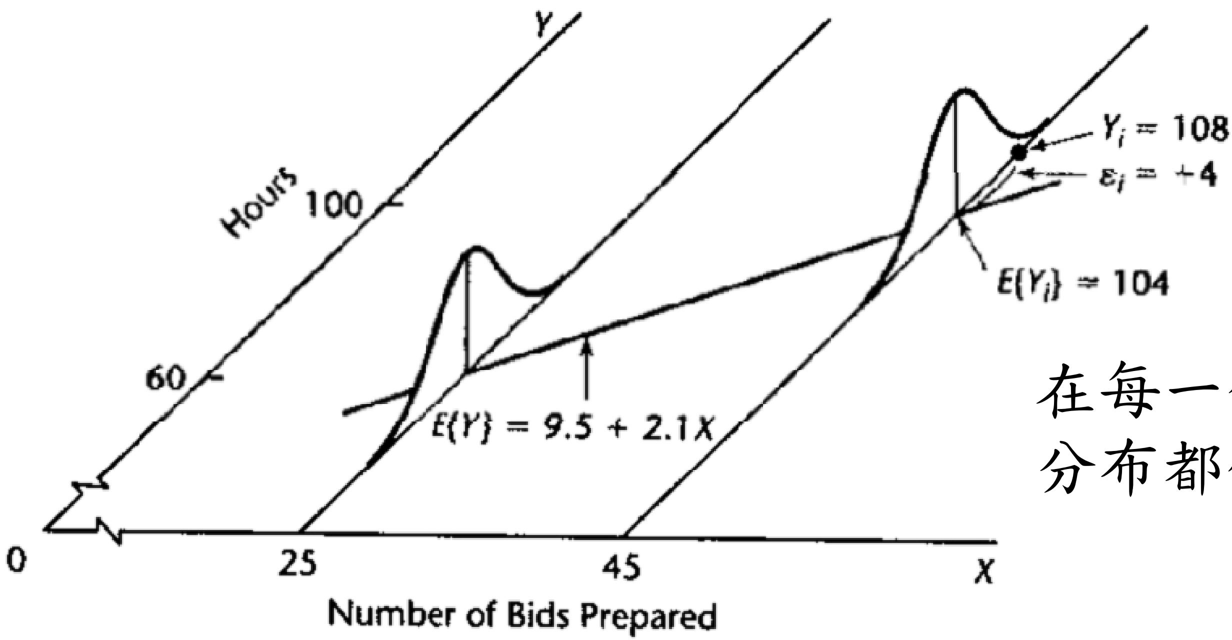
$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_{p-1} X_{1,p-1} + \varepsilon_1$$

$$Y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_{p-1} X_{2,p-1} + \varepsilon_2$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$Y_n = \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_{p-1} X_{n,p-1} + \varepsilon_n$$

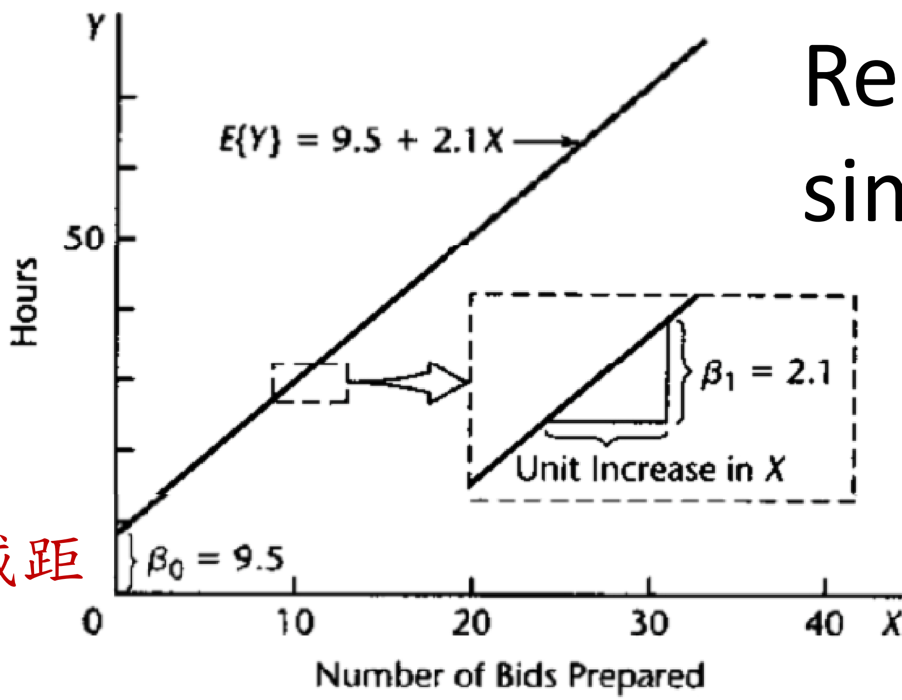
$$\underset{n \times 1}{\mathbf{Y}} = \underset{n \times p}{\mathbf{X}} \underset{p \times 1}{\boldsymbol{\beta}} + \underset{n \times 1}{\boldsymbol{\varepsilon}}$$



$$\varepsilon_i \sim N\left(0, \sigma^2\right)$$

在每一個X值之下，Y的分布都假設為常態分布

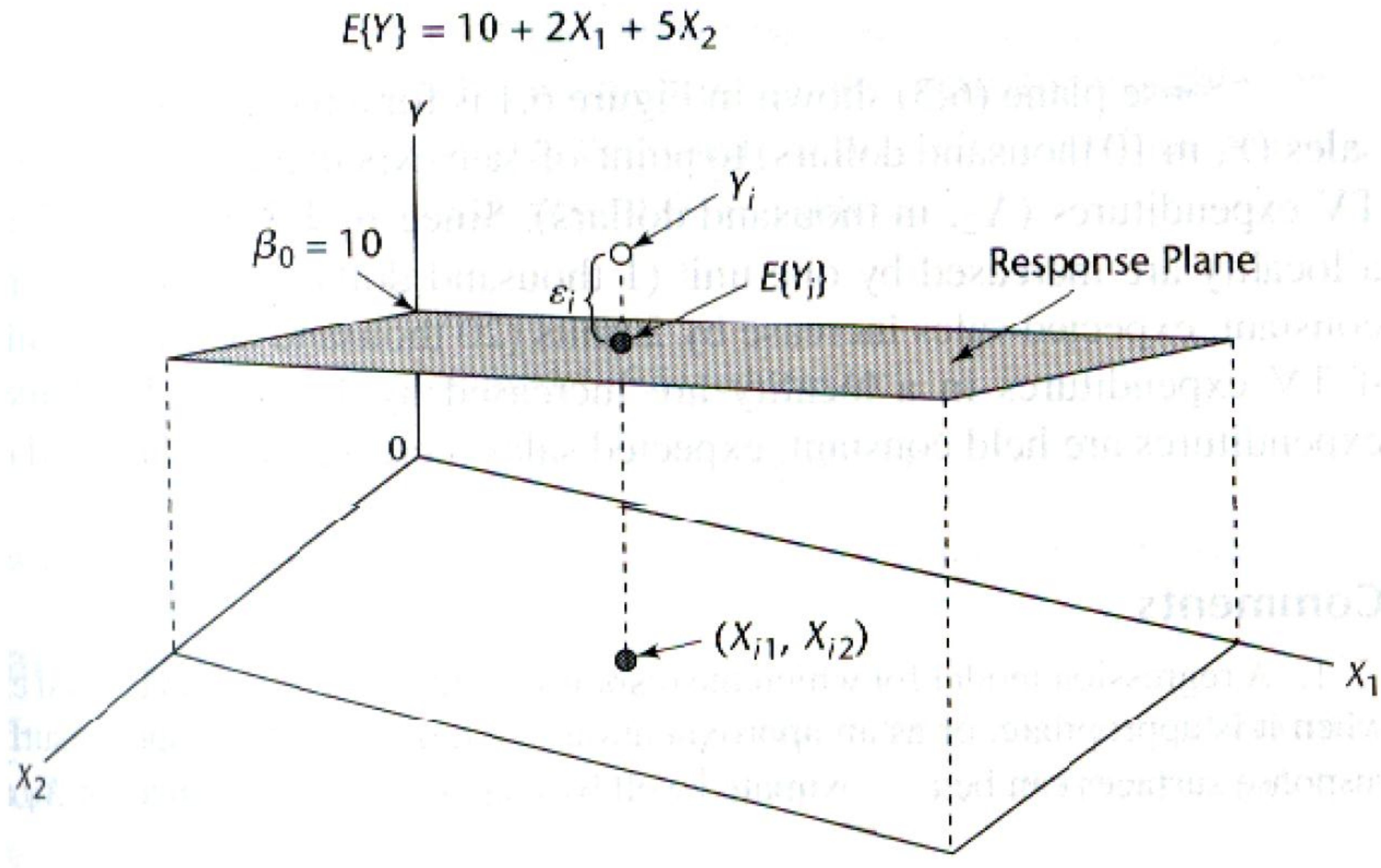
Regression coefficients in simple linear regression



截距

$\hat{\beta}_1$：每增加一單位的X，Y平均增加 2.1 個單位 (斜率)

Regression coefficients in multiple linear regression



Regression coefficients in multiple linear regression

- $\hat{\beta}_1$：在相同的X_2下，每增加一單位的X_1，Y平均增加 2 個單位
- $\hat{\beta}_2$：在相同的X_1下，每增加一單位的X_2，Y平均增加 5 個單位

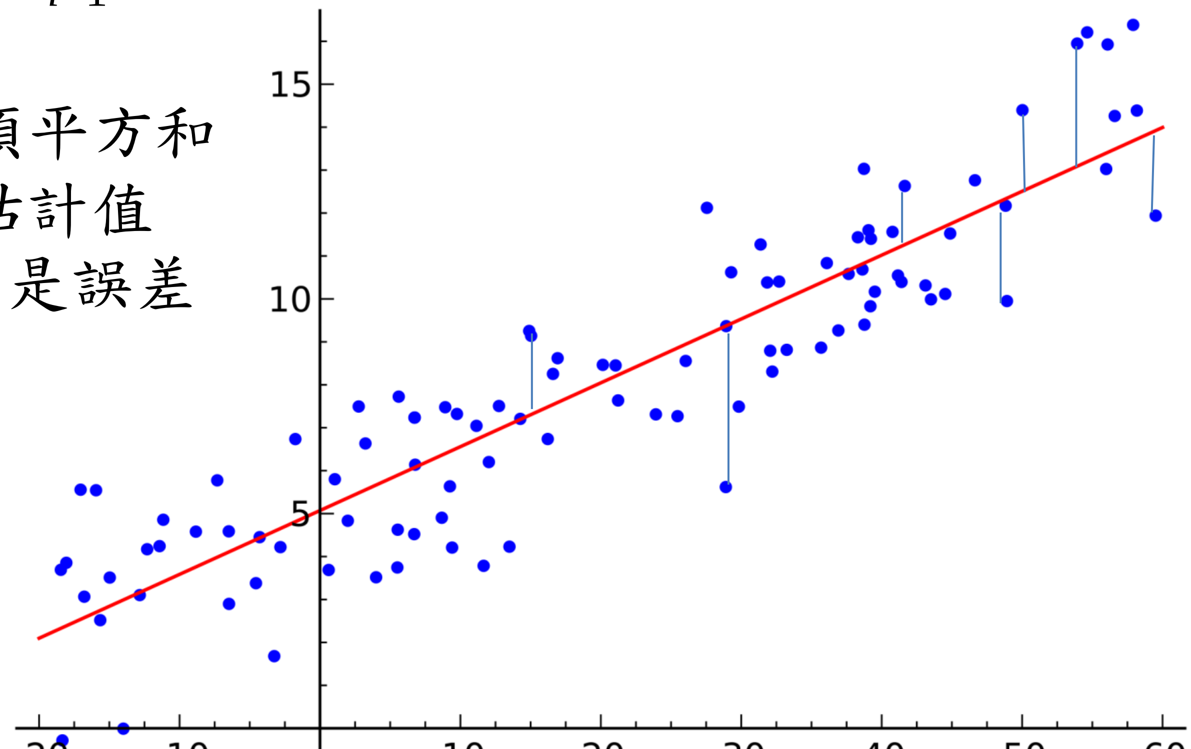
Estimation of regression coefficients

- 最小平方法 (method of least squares)

$$\min \sum_{i=1}^{n} \varepsilon_i^2 = \min \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i\right)^2$$

想法：尋找能使誤差項平方和達到最小的迴歸係數估計值

Q：為什麼目標函式不是誤差項和？



r：X和Y之間的相關係數 (correlation coefficient)

$$-1 \le r \le 1$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = r\frac{S_y}{S_x}$$

$$= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \cdot \frac{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \Big/ n-1}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \Big/ n-1}}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Correlation does not imply causation (因果關係).

A statistically significant regression coefficient does not imply causation.

迴歸係數估計值的矩陣表示法

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

可以用SAS或R等統計軟體求得

其中 $\boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1,p-1} \\ 1 & X_{21} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{n,p-1} \end{bmatrix}, \quad \boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$

RGui (64-bit) - [R Console]

檔案 編輯 看 其他 程式套件 視窗 輔助

```
>
> bb <- read.csv('I:/AppliedBiostatistics/LinWY/BMI.csv')
> rr <- lm(bb$BMI ~ bb$SEX + bb$AGE + bb$rs35682)
> summary(rr)

Call:
lm(formula = bb$BMI ~ bb$SEX + bb$AGE + bb$rs35682)

Residuals:
    Min      1Q  Median      3Q     Max
-11.042  -4.529  -1.754   3.116  23.276

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.86616    2.97906   4.319 3.82e-05 ***
bb$SEXM      0.29542    1.54432   0.191   0.8487
bb$AGE       0.28282    0.06313   4.480 2.06e-05 ***
bb$rs35682   1.94442    0.90597   2.146   0.0344 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.246 on 96 degrees of freedom
Multiple R-squared: 0.206,	Adjusted R-squared: 0.1811
F-statistic:   8.3 on 3 and 96 DF,  p-value: 5.783e-05

>
>
>
>
```

ID	SEX	AGE	BMI	rs35682
ID1	F	35	24	0
ID2	M	32	24	1
ID3	F	30	21	1
ID4	F	31	21	0
ID5	F	52	22	1
ID6	F	59	22	0
ID7	M	57	44	0
ID8	F	34	21	1

BMI data set

統計軟體 R

好處：免費、且有很多最新方法學發展出來的packages開放供大眾下載使用

RGui (64-bit) - [R Console]

檔案 編輯 看 其他 程式套件 視窗 輔助

```
> bb <- read.csv('I:/AppliedBiostatistics/LinWY/BMI.csv')
> Y <- bb$BMI
> for(i in 1:nrow(bb)){
+   if(bb$SEX[i]=='M'){
+     bb$SEX01[i] <- 1
+   }
+   if(bb$SEX[i]=='F'){
+     bb$SEX01[i] <- 0
+   }
+ }
> X <- cbind(rep(1,nrow(bb)), bb$SEX01, bb$AGE, bb$rs35682)
> (solve(t(X)%*%X))%*%(t(X)%*%Y)
           [,1]
[1,] 12.8661591
[2,]  0.2954208
[3,]  0.2828177
[4,]  1.9444206
>
```

寫迴圈

流行病學與生物統計計算（下學期選修課）

Computing in Epidemiology and Biostatistics

授課教師：林菀俞；李文宗

迴歸係數的解釋

- 在相同的性別與年紀下，rs35682 上 ‘A’ allele 的個數每增加一個，BMI 平均會增加 1.944 kg/m^2

Nominal variable（名目變項）：性別（男、女）、血型 (A 、 B 、 O 、 AB)

Ordinal variable（序位變項）：用功程度分類（都不唸書、考試前才唸書、平常就有在唸書、一直都很用功唸書）

Interval variable（等距變項）：氣溫（未必有絕對的零點）

Ratio variable（等比變項）：身高、體重（有絕對的零點）

Variance-covariance matrix of regression coefficients

$$\operatorname{var}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

Mean squared error (MSE)

$$\widehat{\operatorname{var}}\left(\hat{\boldsymbol{\beta}}\right) = \hat{\sigma}^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} = \mathrm{MSE} \cdot \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

ε_i ： 誤差(error)

e_i ： 殘差(residual)

$$\mathrm{MSE} = \hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2$$

$$= \frac{1}{n-p}\sum_{i=1}^{n}\left[Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_{p-1} X_{i,p-1}\right)\right]^2$$

$$= \frac{1}{n-p}\boldsymbol{e}'\boldsymbol{e} = \frac{1}{n-p}\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)$$

Variance-covariance matrix of regression coefficients

$$\hat{\text{var}}\left(\hat{\boldsymbol{\beta}}\right) = \hat{\sigma}^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} = \text{MSE} \cdot \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

Variable	Intercept	sex01	AGE	rs35682
Intercept	8.8748	0.1033	-0.1726	-0.8723
sex01	0.1033	2.3849	-0.0197	-0.1032
AGE	-0.1726	-0.0197	0.0040	0.0042
rs35682	-0.8723	-0.1032	0.0042	0.8208

Testing the regression coefficient

- With consideration of sex and age, is SNP rs35682 a statistically significant explanatory variable for BMI (given significance level of 5%)?

$$H_0 : \beta_3 = 0 \quad vs. \quad H_1 : \beta_3 \neq 0$$

$$t-value = \frac{\hat{\beta}_3 - \beta_3}{s.e.\left(\hat{\beta}_3\right)} = \frac{\hat{\beta}_3 - \beta_3}{\sqrt{\hat{\mathrm{var}}\left(\hat{\beta}_3\right)}} = \frac{1.94442 - 0}{\sqrt{0.8208}}$$

$$= 2.146 > t_{0.975;96} = 1.985$$

Two-tailed test

$$n - p = 96$$

With consideration of sex and age, SNP rs35682 is a statistically significant explanatory variable for BMI (given significance level of 5%).
在考慮性別與年齡下， SNP rs35682對BMI而言是個統計上顯著的解釋因子(給定顯著水準為0.05時)。

Interval estimation of the regression coefficient

$(1-\alpha)\times 100\%$ *confidence interval for* $\hat{\beta}_j$ $\qquad \hat{\beta}_j \mp t_{1-\alpha/2;n-p} \times s.e.\left(\hat{\beta}_j\right)$

95% *confidence interval for* $\hat{\beta}_3$

$$\hat{\beta}_3 \mp t_{0.975;96} \times s.e.\left(\hat{\beta}_3\right)$$

$$= 1.94442 \mp 1.985 \times \sqrt{0.8208}$$

$$= \left[0.146,\ 3.743\right]$$

SAS or R output

Degrees of freedom （自由度）

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	12.86616	2.97906	4.32	<.0001
sex01	1	0.29542	1.54432	0.19	0.8487
AGE	1	0.28282	0.06313	4.48	<.0001
rs35682	1	1.94442	0.90597	2.15	0.0344

$$s.e.\left(\hat{\beta}_3\right) = \sqrt{\hat{\mathrm{var}}\left(\hat{\beta}_3\right)} = \sqrt{0.8208}$$

$\hat{\beta}_3$

ANOVA table
(Analysis of Variance table)

Source of variation		DF	Sum of Squares	Mean Square	F Value	Pr > F
Regression	$p-1=$	3	1307.40520	435.80173	8.30	<.0001
Error	$n-p=$	96	5040.70480	52.50734		
Total	$n-1=$	99	6348.11000			

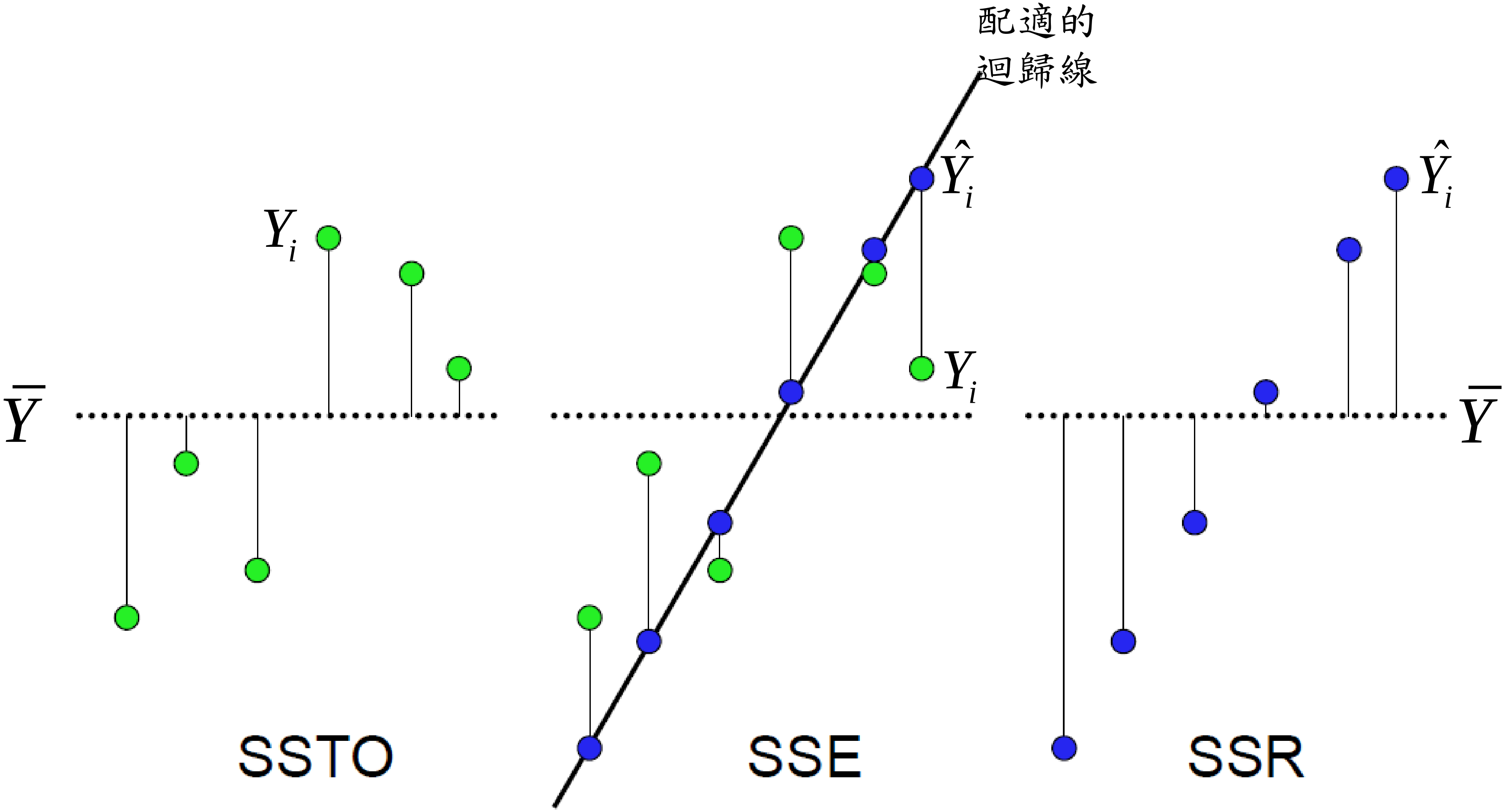
Regression sum of squares $$SSR=\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2,\quad MSR=\frac{SSR}{p-1}$$

Error sum of squares $$SSE=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=\sum_{i=1}^{n}e_i^2,\quad MSE=\frac{SSE}{n-p}$$

Total sum of squares $$SSTO=\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2=\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2+\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=SSR+SSE$$

$$F-value=\frac{MSR}{MSE}=\frac{435.8}{52.5}=8.3>F_{0.95;3,96}=2.699$$

Sum of Squares (平方和)



$$\mathrm{SSTO} = \mathrm{SSE} + \mathrm{SSR}$$

原始圖摘錄自李文宗教授之投影片

Linear regression model 正式寫法

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

- 其中 Y 為身體質量指數(BMI)，X_1 為性別，X_2 為年齡，X_3 為 rs35682上‘A’對偶基因的個數
- $i = 1, 2, \cdots, n$ 代表個案指標
- 誤差項 ε_i $(i = 1, 2, \cdots, n)$ 假設為彼此獨立的常態分布，服從 $N\left(0, \sigma^2\right)$

檢定整體迴歸關係是否存在

- 命題：檢定BMI 與性別、年齡和rs35682上‘A’ 對偶基因個數的整體迴歸關係是否存在？

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0 \quad vs.$$

$$H_1: at\ least\ a\ \beta_j \neq 0\ \ (j = 1,2,3)$$

$$F - value = \frac{MSR}{MSE} = \frac{435.8}{52.5} = 8.3 > F_{0.95;3,96} = 2.699$$

- 結論：給定顯著水準為0.05， BMI 與性別、年齡和rs35682上‘A’對偶基因個數的整體迴歸關係具有統計上的顯著意義

Coefficient of multiple determination （複判定係數）

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \qquad 0 \le R^2 \le 1$$

$$R^2 = \frac{1307.40520}{6348.11000} = 0.206$$

- 性別、年齡與rs35682上‘A’對偶基因的個數可以解釋身體質量指數的變異達20.6%
- 每增加一個解釋變數，R^2值都會上升，不管該變數是否真具有統計上顯著的解釋意義

Adjusted coefficient of multiple determination （調整後的複判定係數）

$$R_a^2 = 1 - \frac{\frac{SSE}{n-p}}{\frac{SSTO}{n-1}} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SSTO}$$

- 不同於複判定係數，在加入一個不甚具有解釋力的自變項時，R_a^2有可能降低
- 思考：為什麼？

檢定部分的自變項是否可解釋Y

- 命題：把性別納入考量後，檢定年齡和rs35682上‘A’對偶基因個數是否有額外解釋BMI的能力？

性別　年齡　rs35682

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

$$H_0 : \beta_2 = \beta_3 = 0 \quad vs. \quad H_1 : at\ least\ a\ \beta_j \neq 0\ \ (j = 2,3)$$

- Full model: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$
- Reduced model: $Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$

Full model

Source of variation	DF	Sum of Squares	Mean Square	F Value	Pr > F
Regression	3	1307.40520	435.80173	8.30	<.0001
Error	$\boxed{96}$	$\boxed{5040.70480}$	52.50734		
Total	99	6348.11000			

df_F SSE_F

Reduced model

Source of variation	DF	Sum of Squares	Mean Square	F Value	Pr > F
Regression	1	79.80444	79.80444	1.25	0.2667
Error	$\boxed{98}$	$\boxed{6268.30556}$	63.96230		
Total	99	6348.11000			

df_R SSE_R

Partial F test (偏F檢定)

$$F-value=\frac{\dfrac{SSE_R-SSE_F}{df_R-df_F}}{\dfrac{SSE_F}{df_F}}=\frac{MSR(X_2,X_3\mid X_1)}{MSE(X_1,X_2,X_3)}\sim F_{df_R-df_F,\ df_F}$$

在本例中 $F-value=\dfrac{\dfrac{6268.30556-5040.70480}{98-96}}{\dfrac{5040.70480}{96}}$

$=11.69>F_{0.95;2,96}=3.091$

把性別納入考量後，年齡和rs35682上‘A’ 對偶基因個數的確有額外解釋BMI的能力(顯著水準5%)

Coefficient of partial determination (偏判定係數)

- 在本例中

$$R^2_{Y,X_2,X_3|X_1} = \frac{SSE_R - SSE_F}{SSE_R} = \frac{SSE(X_1) - SSE(X_1, X_2, X_3)}{SSE(X_1)}$$

$$= \frac{6268.30556 - 5040.70480}{6268.30556} = 0.1958$$

把性別納入考量後，年齡和rs35682上‘A’對偶基因個數可以額外解釋身體質量指數的變異達19.58%

Homework 2012.10.23

- (BMI data set) With consideration of sex and age, is SNP rs35682 a statistically significant explanatory variable for BMI (given significance level of 5%)? Please note that a linear trend between the number of allele 'A' and BMI cannot be assumed.

  (1) Linear regression model

  (2) Matrix presentation for $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} \boldsymbol{X}'\boldsymbol{Y}$

  (3) Statistical inference (Partial F test, Coefficient of partial determination, the interpretation for each estimated regression coefficient)

提示：dummy variable

先recode變數

	X_3	X_4
0 (aa)	0	0
1 (Aa)	1	0
2 (AA)	0	1

在相同的性別與年齡下， SNP rs35682基因型為Aa的人比基因型為aa的人BMI平均多XX……
在相同的性別與年齡下， SNP rs35682基因型為AA的人比基因型為aa的人BMI平均多XX……