R 語言與機器學習 (三)

丘祐瑋
David Chiu

迴歸問題
(Regression Analysis)

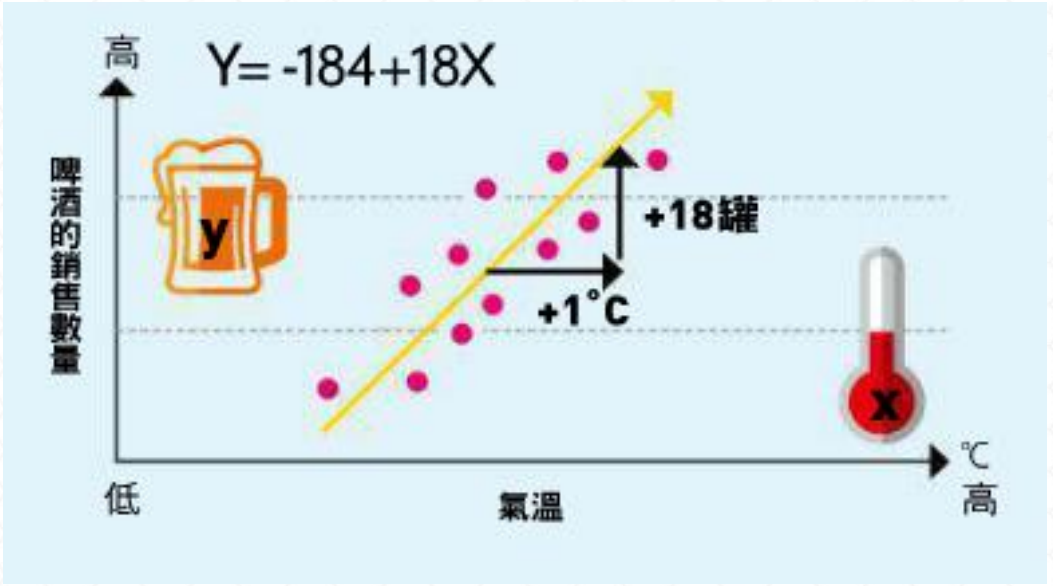
回歸分析

- 線性回歸是研究**單一依變項**(dependent variable)與一個或以上**自變項**(independent variable)之間的關係

- 線性回歸有兩個主要用處：
  - **預測**指的是用已觀察的變數來預測依變項
  - **因果分析**則是將自變項當作是依變項發生的原因
- **Francis Galton 在1886 年發表論文Regression Towards Mediocrity in Hereditary Stature，認為孩子身高跟父親的身高成正相關，而身高的變異數不會隨時間而增加。**

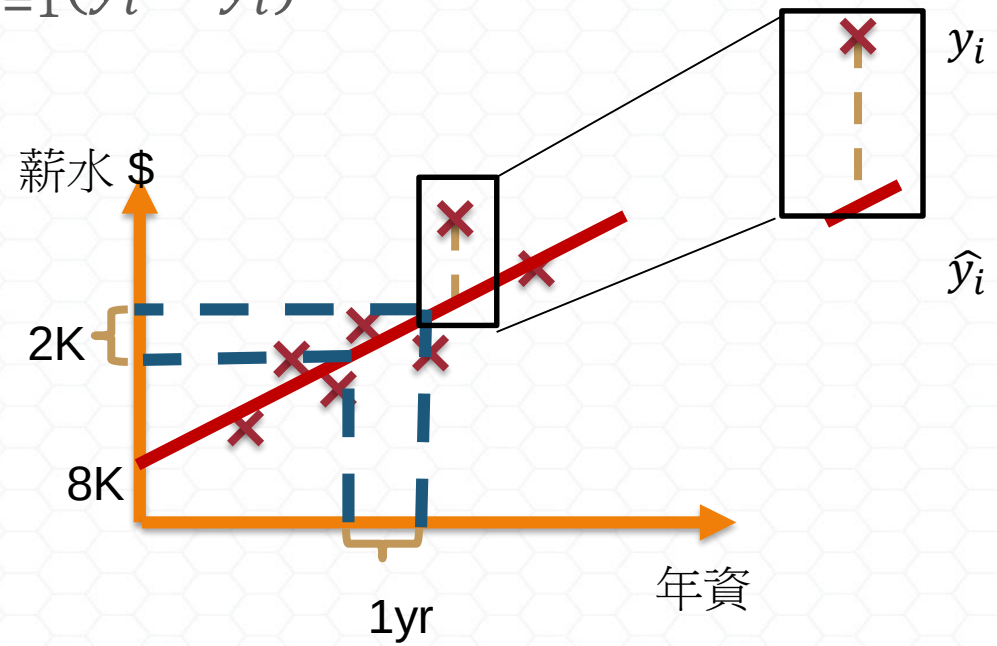
簡單線性迴歸

- 數學模型
  - $y = \beta_1 x + \beta_0 + \epsilon$
  - y 是依變數
  - x 是自變數
  - b_i 是迴歸係數
  - b_0是截距
  - ϵ是誤差



最小平方估計法 - OLS

- 找出殘差平方和最小的一條線
  - 殘差$e_i = (y_i - \hat{y}_i)$
  - 殘差平方和 $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$



計算迴歸係數

X (kilos)	Y (cost, $)
17	132
21	150
35	160
39	162
50	149
65	170

- 計算兩變數之間的迴歸係數

```
> X <- c(17, 21, 35, 39, 50, 65)
> Y <- c(132, 150, 160, 162, 149, 170)
> X_avg <- mean(X)
> Y_avg <- mean(Y)
> B1 <- (sum(X*Y) - 6 * X_avg * Y_avg)/(sum(X^2) - 6 * X_avg ^2)
```

$$\hat{\beta}_1 = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}$$

計算截距

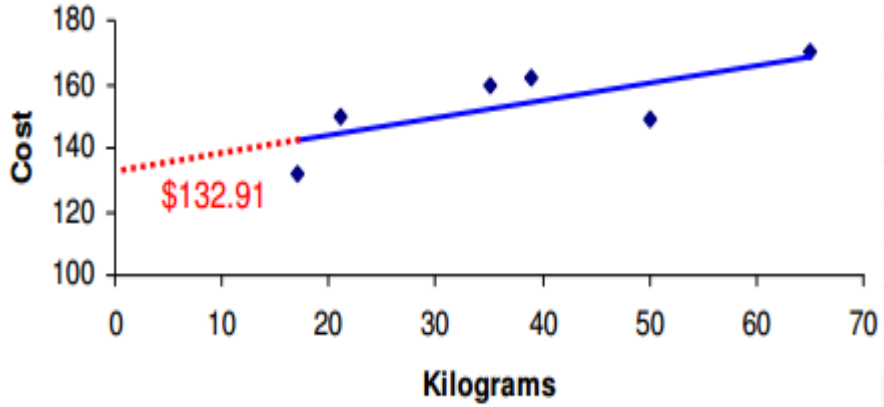
- 計算截距

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

```
> B0 <- Y_avg - B1 * X_avg
```

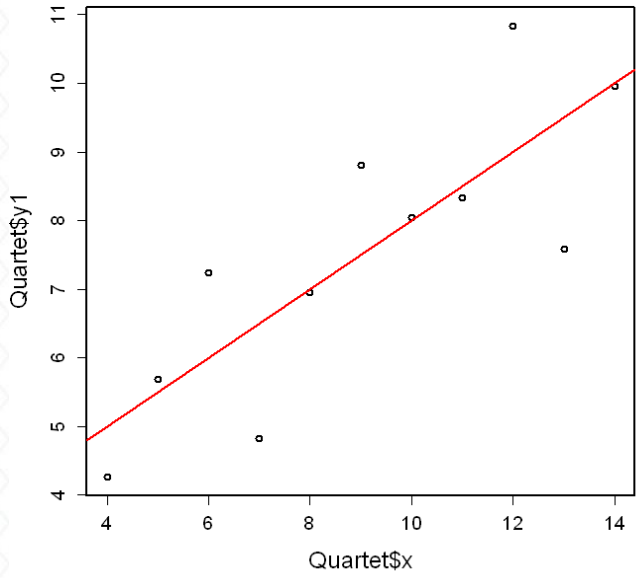
- 得到迴歸公式

  $y = 0.553x + 132.91$



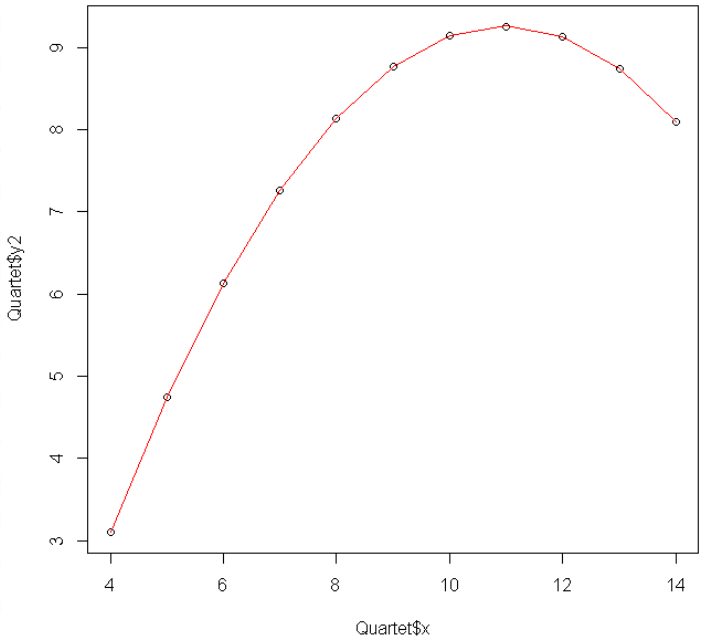
一項式回歸分析

```
plot(y1~x1, data = anscombe)
lmfit <- lm(y1~x1, data = anscombe)
abline(lmfit, col="red")
```



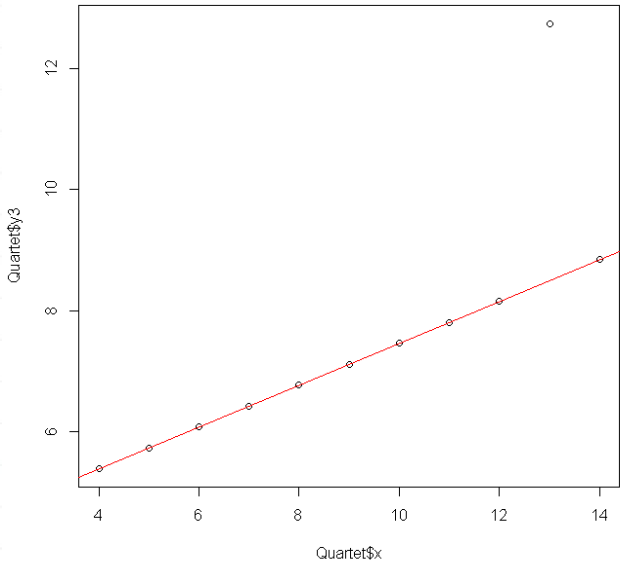
二項式回歸分析

```
plot(y2~x1, data = anscombe)
lmfit <- lm(y2~poly(x1,2), data = anscombe)
lines(sort(anscombe$x), lmfit$fit[order(anscombe$x)], col = "red")
```



可容錯的回歸 (rlm)

```
plot(y3~x1, data = anscombe)
lmfit <- rlm(y3~x1, data = anscombe)
abline(lmfit, col="red")
```

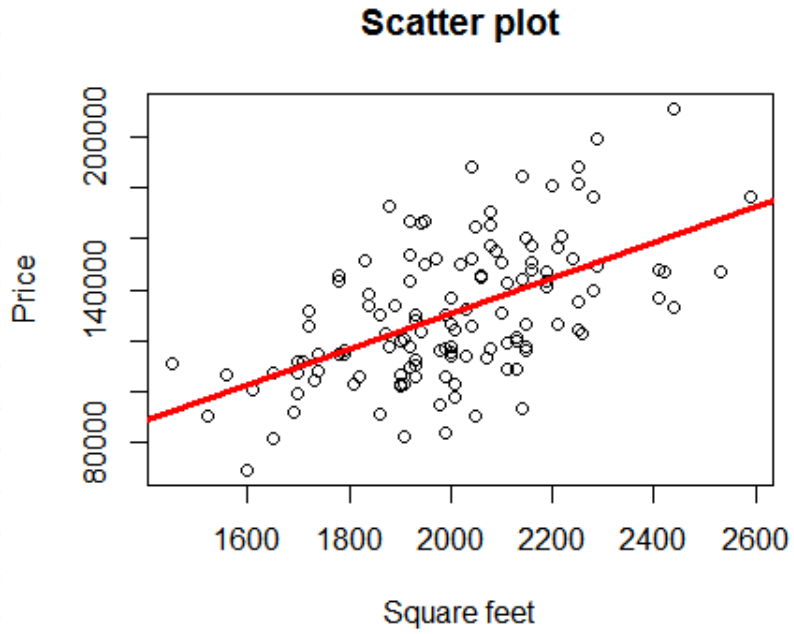


匯入資料

```
library(car)
library(MASS)
house_prices <- read.csv(file="house-prices.csv")
names(house_prices)
str(house_prices)
attach(house_prices)
```

繪出迴歸線

```
lm.1 <- lm(Price ~ SqFt, data=house_prices)
summary(lm.1)
plot(SqFt, Price, main="Scatter plot", xlab="Square feet", ylab="Price")
abline(lm.1,col="red",lwd=3)
```



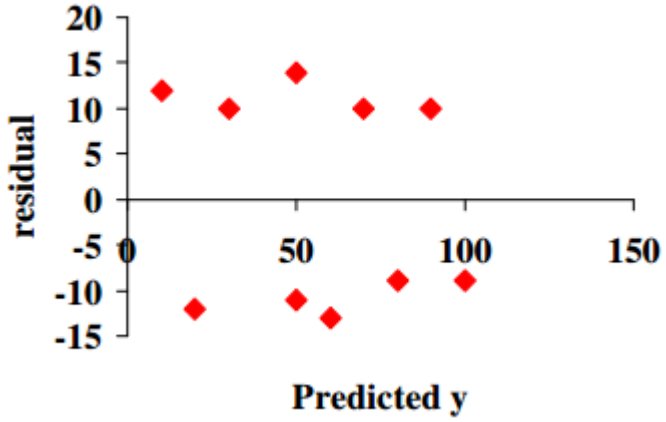
誤差(隨機項)的假設

- 最小平方估計法 OLS 適用於
  - 誤差變數是常態分佈
    - 可以使用Shapiro 檢定,或用直方圖驗證
  - 錯誤變數的期望值是0
    - 使用最小估計法求是否誤差的期望值是0
  - 誤差的變異數是常數
    - 驗證變異數齊一性 (Homoscedasticity)
  - 任兩個依變數所關聯的錯誤互為獨立
    - 驗證$Cov(\varepsilon_1, \varepsilon_2) = 0$

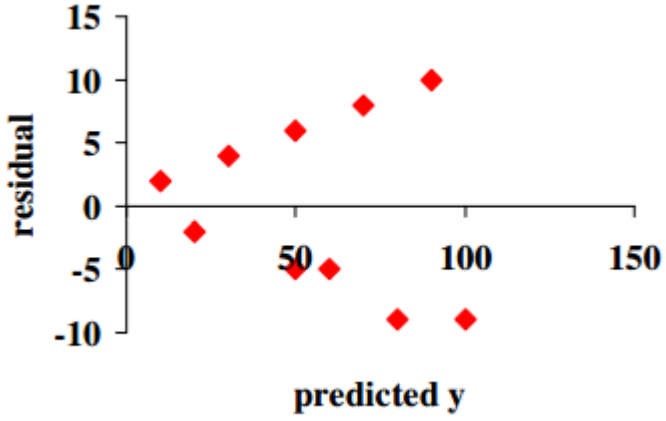
驗證變異數齊一性(Homoscedasticity)

- 齊一性(Homoscedasticity): 增加X 時, Y 會穩定增長
- 相異性(Heteroscedasticity): 增加X 時, Y 增長幅度較大(小)

Homoscedasticity

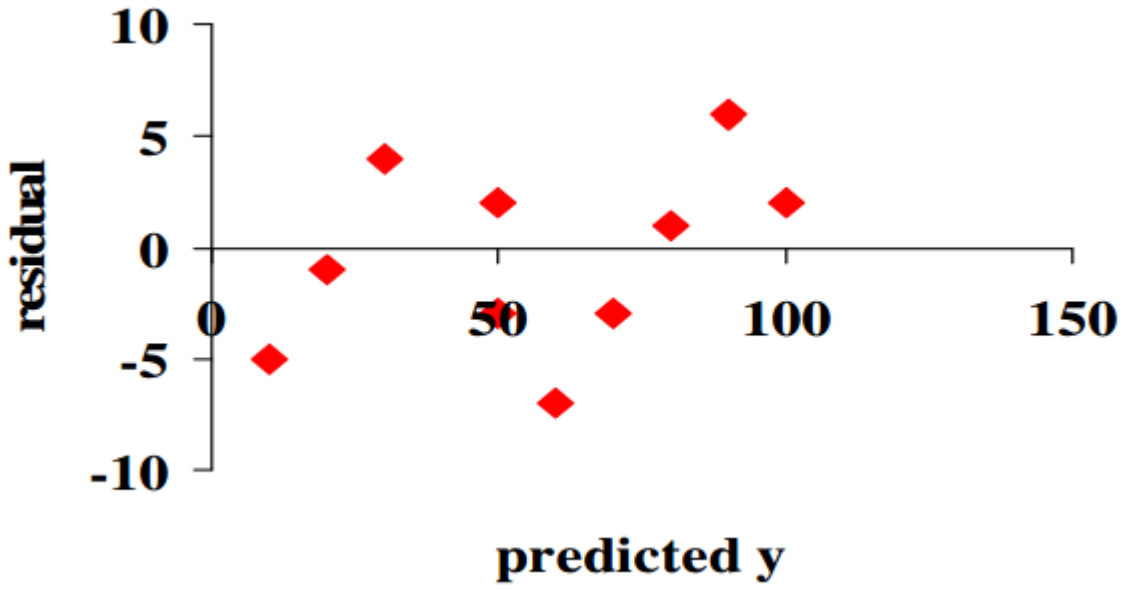


Heteroscedasticity



驗證錯誤是否獨立

- 如果$Cov(\varepsilon_1, \varepsilon_2) \neq 0$ 代表所預測出來的Y 值會受X值產生的順序影響，稱為自我相關性(autocorrelation)。通常發生在時間序列上



驗證迴歸模型

- 線性模型是否合適?
  - 使用散佈圖看是否有線性關係
  - 檢視殘差圖
  - 檢視誤差變數(隨機項)
- 兩變數之間是否真的有線性關係?
- 兩變數相關的程度多強？

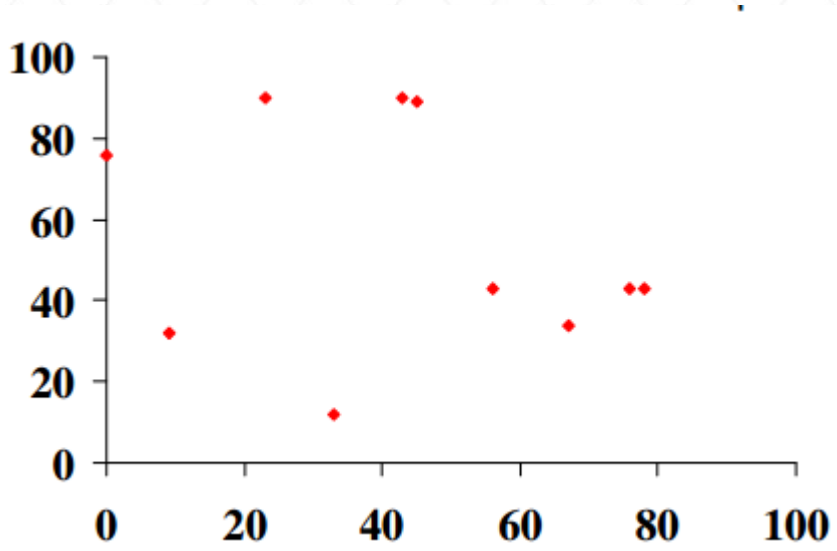
殘差 (Residual)

- 當線性模型配適資料良好時，殘差比較小，變異數也較小
  - 變異數 $S_\varepsilon{}^2$
  - 標準差 S_ε

- 可以使用殘差的標準差(Residual Standard Error)來量測線性模型的適配性，但數值大小端看依變數的取樣數，所以只適合用做模型的比較

線性關係顯著性檢定

- $y = \beta_1 x + \beta_0 + \epsilon$



如資料之間沒有相關性，β_1 應該接近於0，但事實上β_1不太可能等於零，因此要如何驗證 β_1 近似於0?

驗證相關性

- 假設如下:
  - Ho: $\beta_1 = 0$
  - Ha: $\beta_1 \neq 0$
  - test : $$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$$
- 假設虛無假設正確 $$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

參數估計

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	132.9129896	10.10788494	13.14943634	0.000193179	104.8489438	160.9770354
x	0.552960628	0.245140203	2.255691322	0.087094659	-0.127659098	1.233580354

迴歸係數

標準差

當虛無假設成立時的t

t 發生的機率

95%信心水準估計

參數估計(續)

- 假設顯著性標準是0.01
- 推翻虛無假設的標準是 p 值< 0.01

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	132.9129896	10.10788494	13.14943634	0.000193179	104.8489438	160.9770354
x	0.552960628	0.245140203	2.255691322	0.087094659	-0.127659098	1.233580354

- t = 2.26, P(>t) =0.087, $t_{0.01}$=2.62
- **驗證兩者關係並非顯著**

R Square

- 為相關係數(correlation)的平方值，其值介於0與1之間
  - R Square越接近1，表示自變數與依變數相關性越高
  - 越接近0，表示自變數與依變數相關性越低
  - 迴歸可以解釋的變異比例，可作為自變數預測依變數準確度的指標

模型評估

- Relative Mean Square Error
  - 評估使用同單位的模型
  - $$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}}$$
- Relative Square Error
  - 評估使用不同單位的模型
  - $$RSE = \frac{\sum_{i=1}^{n}(p_i - a_i)^2}{\sum_{i=1}^{n}(\bar{a} - a_i)^2}$$

R Square

- 迴歸可以解釋的變異比例，可作為自變數預測依變數準確度的指標

Coefficient of Determination → $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

Sum of Squares Total → $SST = \sum (y - \bar{y})^2$

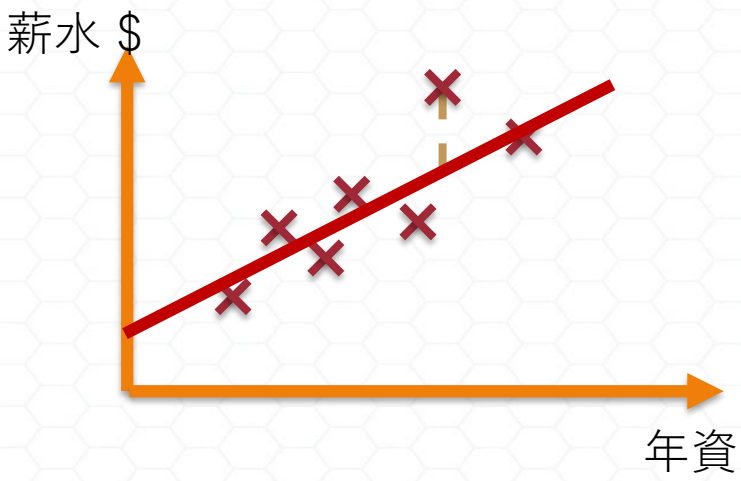
Sum of Squares Regression → $SSR = \sum (y' - \bar{y'})^2$

Sum of Squares Error → $SSE = \sum (y - y')^2$

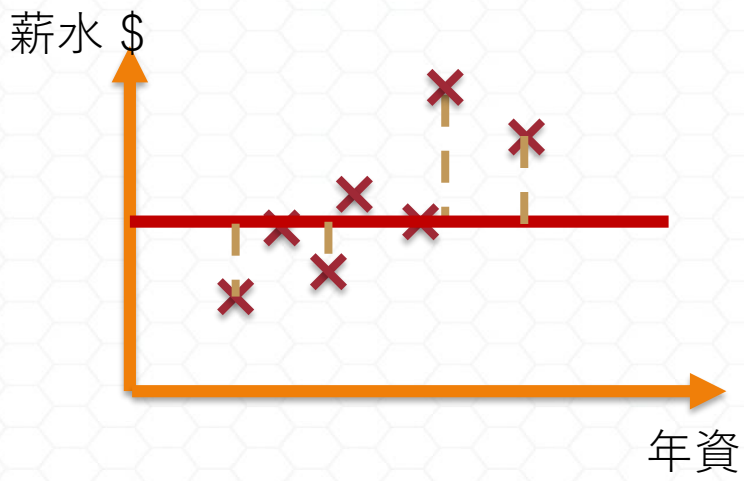
R Squared

- 回歸可以解釋的變異比例，可作為引數預測依變數準確度的指標

殘差平方和 $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$



整體平方和 $SST = \sum_{i=1}^{n}(y_i - y_{avg})^2$



$$R^2 = 1 - \frac{SSE}{SST}$$

評估一般線性模型

```
plot(y3 ~ x1, data = anscombe)
lmfit <- lm(y3~x1, data = anscombe)
abline(lmfit, col="red")
predicted <- predict(lmfit, newdata=anscombe[c("x1")])
actual <- anscombe$y3
rmse <- (mean((predicted - actual)^2))^0.5
mu <- mean(actual)
rse <- mean((predicted - actual)^2) / mean((mu - actual)^2)
rsquare <- 1 - rse
```

評估可容錯的模型

```
library(MASS)
plot(y3~x1, data = anscombe)
rlmfit <- rlm(y3~x1, data = anscombe)
abline(rlmfit, col="red")
predicted = predict(rlmfit, newdata=anscombe[c("x")])
actual <- anscombe$y3
rmse <- (mean((predicted - actual)^2))^0.5
mu <- mean(actual)
rse <- mean((predicted - actual)^2) / mean((mu - actual)^2)
rsquare <- 1 - rse
```

模型驗證總結

- 檢視 R square > 0.7
- 檢視Multicollinearity
  - vif 是否 < 10
  - vif: 評估迴歸係數的變異數是否因共線性而增加
- 檢視 P values 是否 < .05
- 檢視殘差圖

多元迴歸分析 (Multiple Regression)

- 當要探索多個自變量與一個依變量之間的關係時
  - e.g. Sales = 100+25print -100TV+67Radio

- 基礎假設
  - 依變量為隨機變數
  - 變數之間有統計關聯
  - 自變數與依變數有線性關係
  - 自變數之間的共線性(Co-linearity 最小)

數學模型

- $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots \beta_k x_{ki} + \beta_0 + \epsilon$
  - $x_1 \cdots x_k$ 為自變量
  - y 是依變數
  - b_i 是迴歸係數
  - b_0是截距
  - ϵ是誤差
- 目標:
  - 最小化殘差平方和

$$\mathrm{SSR} = \sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

建立變數

```
house_prices$brick_d<-ifelse(house_prices$Brick=="Yes",1,0)
house_prices$east<-
ifelse(house_prices$Neighborhood=="East",1,0)
house_prices$north<-
ifelse(house_prices$Neighborhood=="North",1,0)
```

建立訓練與測試資料集

```
set.seed(110)
sub <- sample(nrow(house_prices), floor(nrow(house_prices) * 0.6))
training_data <- house_prices[sub,]
validation_data <- house_prices[-sub,]
```

建立多元迴歸模型

```
lm.fit1 <- lm(Price ~ SqFt+Bathrooms+Bedrooms+Offers+
              north+east+brick_d, data=training_data)
summary(lm.fit1)
```

房租預測

- 資料集

Rent, $	360	1000	450	525	350	300
No. of rooms	2	6	3	4	2	1
Distance from Downtown (in miles)	1	1	2	3	10	4

- 參數估計

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standar-dized Estimate	Variance Inflation
Intercept	Intercept	1	96.458	118.12	0.82	0.47	0	0
Number_of_rooms	Number_of_rooms	1	136.48	26.864	5.08	0.01	0.94297	1.23
dis_downtown	Distance_from_Downtown	1	-2.4035	14.171	-0.17	0.88	-0.0315	1.23

- 公式

$$\text{Rent} = 96.458 + (136.48)\text{No. of rooms} + (-2.4035)\text{Distance}$$

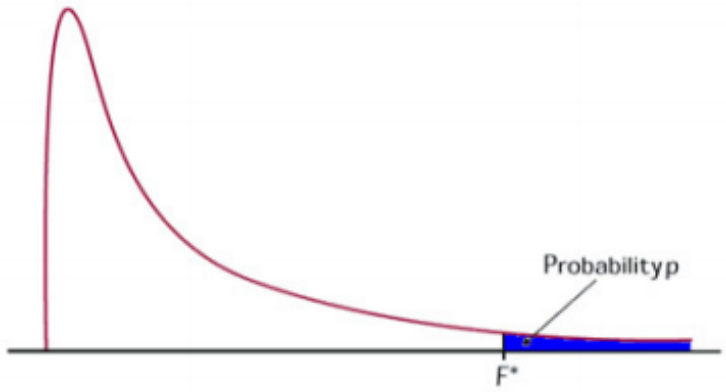
檢驗多元迴歸模型

- 如同檢驗簡單迴歸模型
  - 檢視模型是否適配資料
  - 檢驗每個自變數是否都為顯著
- 檢驗假說

$$H_O : \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$$

$$H_A : \text{at least one } \beta_k \neq 0$$

- 使用F統計



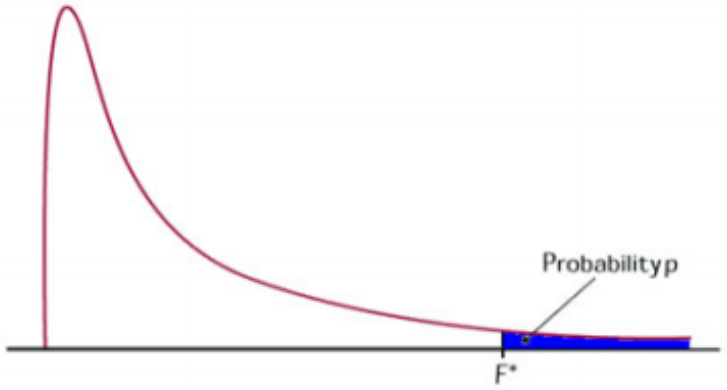
計算F統計

- 迴歸平方和(regression SS) – 依變數的變化歸咎於迴歸模型
- 誤差平方和(error SS) – $\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$ 變化非歸咎於線性模型
- 總和平方和 (total SS) – $\sum_{i=1}^{n}\left(Y_i - \hat{Y}\right)^2$ 體變化

$$\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}\right)^2$$

計算F統計 (續)

- 迴歸平方平均(Model Mean Square)
  - Regression SS / Regression d.f (k)
  - k = 自變數的數量
- 誤差平方平均(Error Mean Square)
  - Error SS / Error d.f. (n-k-1)
  - n = 觀測值的數量
- F統計(F-Statistic)
  - $F = \frac{Model\ Mean\ Square}{Error\ Mean\ Square}$



檢驗模型與變數

- 驗證模型
  - H_0沒有變數顯著
  - H_a至少有一個自變數顯著

- 驗證變數
  - H_0迴歸係數等於0

HO: $\beta_j = 0$

  - H_a迴歸係數不等於0

HA: $\beta_j > 0 \text{ or } \beta_j < 0 \text{ or } \beta_j \neq 0$

其他驗證

- 相關強度

$$\text{Adj.R}^2 = \text{R}^2 - \left[k \frac{\left(1 - \text{R}^2\right)}{n - k - 1} \right]$$

- 殘值分析 (Residual Analysis)
- 直方圖 (確認為常態分佈)
- 殘值對預測值的分析
- 殘值對時間序列的分析
- 殘值對自變數

Adjusted R Square

$$R^2 = 1 - \frac{SSE}{SST}$$

SSE 最小 $\Rightarrow$ R^2 永遠不會遞減

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots \beta_k x_{ki}$$

增加任何一個變數還是會增加R^2

$$Adj\ R^2 = 1 - (1 - \mathrm{R}^2)\ \frac{n-1}{n-p-1}$$

P = 多少回歸因數
n = 總體大小

AIC & BIC

- The Akaike information criterion (AIC) & The Bayesian information criterion (BIC)

$$AIC = 2k + n\ln(SSE/n)$$

- 其中：*K*是參數的數量，*n*為觀察數，*SSE*為殘差平方和

- AIC鼓勵資料擬合的優良性但是儘量避免出現過度擬合（Overfitting）的情況。所以優先考慮的模型應是AIC值最小的那一個。赤池信息量準則的方法是尋找可以最好地解釋資料但包含最少自由參數的模型

減少不顯著的變數

```
lm.fit1.step <- step(lm.fit1)
summary(lm.fit1.step)
```

預測值

■ 訓練資料

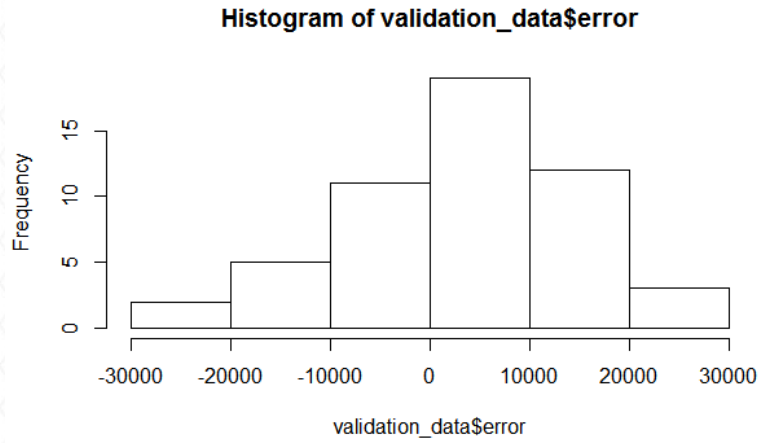
```
training_data$predict.price <- predict(lm.fit1)
training_data$error <- residuals(lm.fit1)
```

■ 測試資料

```
validation_data$predict.price <- predict(lm.fit1,newdata=validation_data)
validation_data$error <- validation_data$predict.price - validation_data$Price
```

檢視殘值

```
hist(training_data$error)
hist(validation_data$error)
```



檢視R Square

```
a<-cor(training_data$Price,training_data$predict.price)
b<-cor(validation_data$Price,validation_data$predict.price)
a*a
b*b
```

分析醫療費用

分析醫療費用

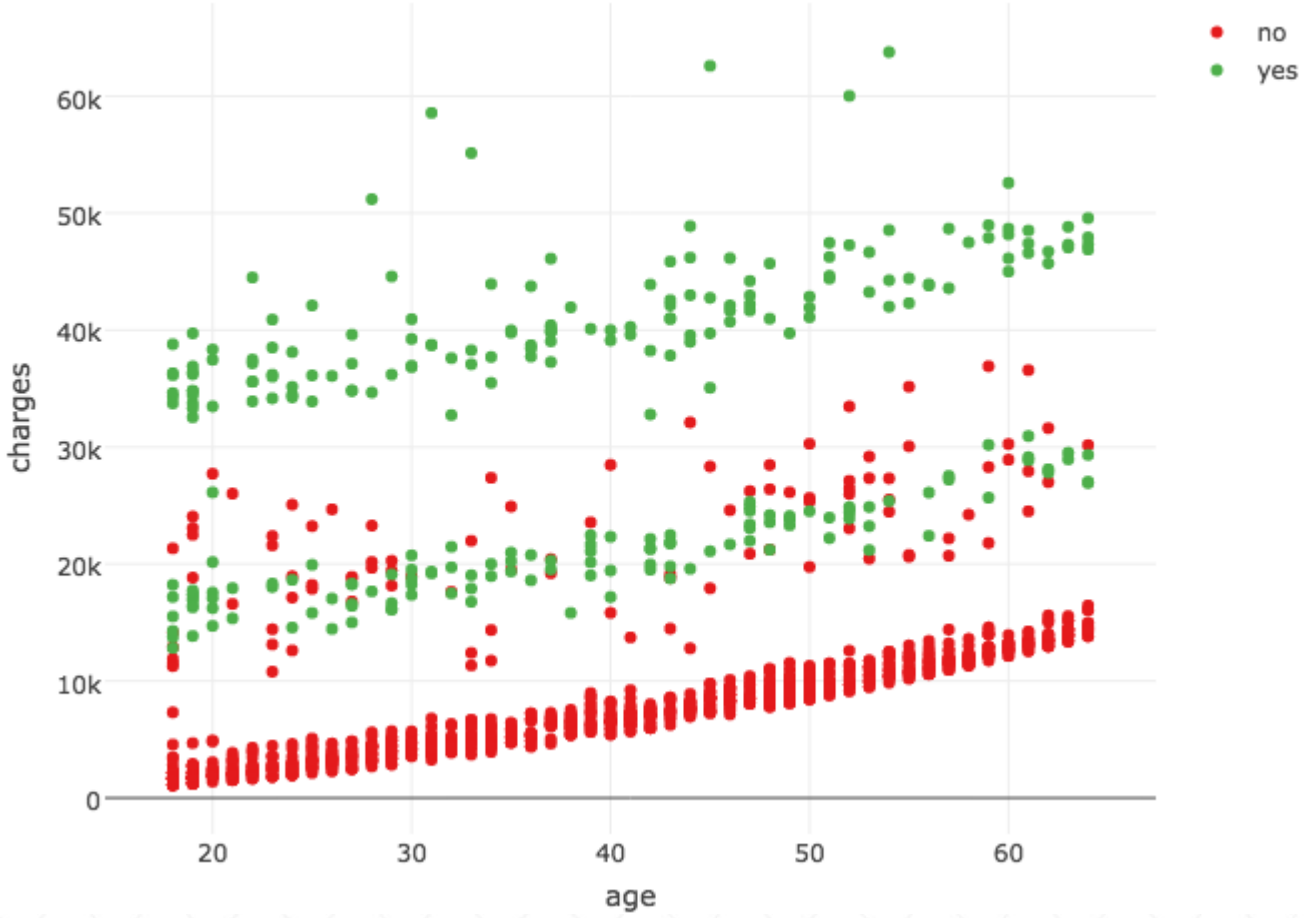
- 預估醫療費用可以幫助保險公司決定該保戶的保險金額，但是如何正確預估醫療費用是件相當困難的事，我們該怎麼使用迴歸分析方法協助保險公司預估每個保護的合理保費呢?

讀取資料集

```
insurance <- read.csv("insurance.csv", stringsAsFactors = TRUE)
class(insurance)
str(insurance)
head(insurance)
```

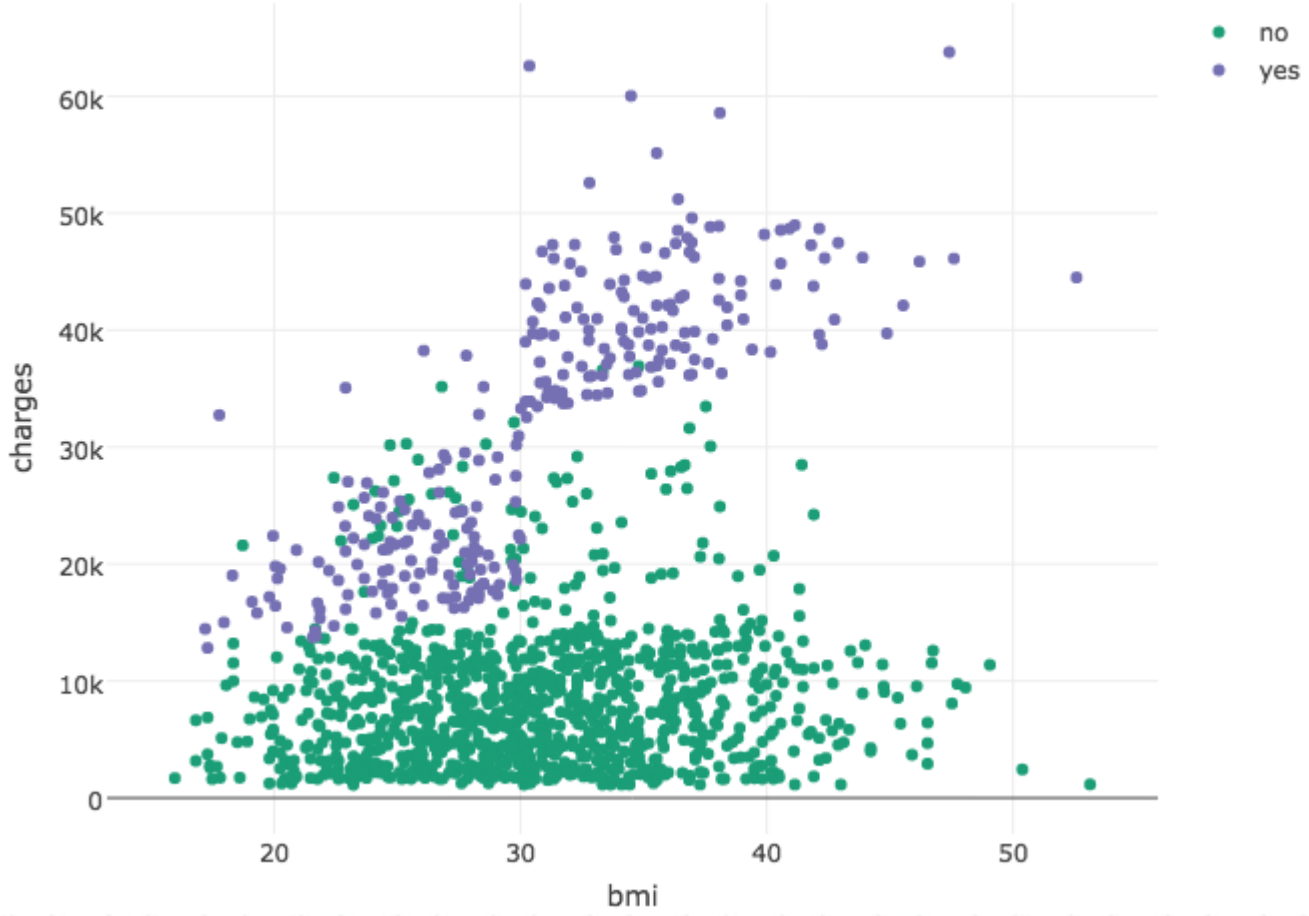
age	sex	bmi	children	smoker	region	charges
19	female	27.900	0	yes	southwest	16884.924
18	male	33.770	1	no	southeast	1725.552
28	male	33.000	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.471
32	male	28.880	0	no	northwest	3866.855
31	female	25.740	0	no	southeast	3756.622

探索資料



有抽菸者，醫療費用比較高昂

探索資料(二)



BMI > 30 , 醫療費用比較高昂

建立迴歸模型

```
#建立模型
ins_model <- lm(charges ~. , insurance)

#檢視模型
ins_model
```

```
##
## Call:
## lm(formula = charges ~ ., data = insurance)
##
## Coefficients:
##     (Intercept)              age          sexmale              bmi
##        -11938.5            256.9           -131.3            339.2
##        children        smokeryes  regionnorthwest  regionsoutheast
##           475.5          23848.5           -353.0          -1035.0
## regionsouthwest
##           -960.1
```

評估模型

summary(ins_model)

```
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
## age                256.9       11.9  21.587  < 2e-16 ***
## sexmale           -131.3      332.9  -0.394 0.693348
## bmi                339.2       28.6  11.860  < 2e-16 ***
## children           475.5      137.8   3.451 0.000577 ***
## smokeryes        23848.5      413.1  57.723  < 2e-16 ***
## regionnorthwest   -353.0      476.3  -0.741 0.458769
## regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
## regionsouthwest   -960.0      477.9  -2.009 0.044765 *
```

增加非線性項

- 不一定所有的變數都跟醫療費用呈線性關係, 例如年紀越大,醫療費用可能倍增

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

```
insurance$age2 <- insurance$age^2
```

將數值資料轉變為類別資料

- 有些資料並非累加性的, 例如: BMI 超過30 可能代表醫療費用較高, 但是 BMI 是 32 抑或是 40 並無明顯差異, 此時我們可以將BMI 轉變為類別資料

```
insurance$bmi30 <- ifelse(insurance$bmi >= 30, 1, 0)
```

交互作用

- 有些變數的交互作用可能比單一指標來的有用, 例如: 抽菸又肥胖的人他的健康狀況應該比單一指標來的更具代表性

bmi30*smoker

重新調整模型

```
ins_model2 <- lm(charges ~ age + age2 + children + bmi + sex +
bmi30*smoker + region, data = insurance)

summary(ins_model2)
```

THANK YOU