Evaluating Regression Models Performance

一、R-Squared Intuition

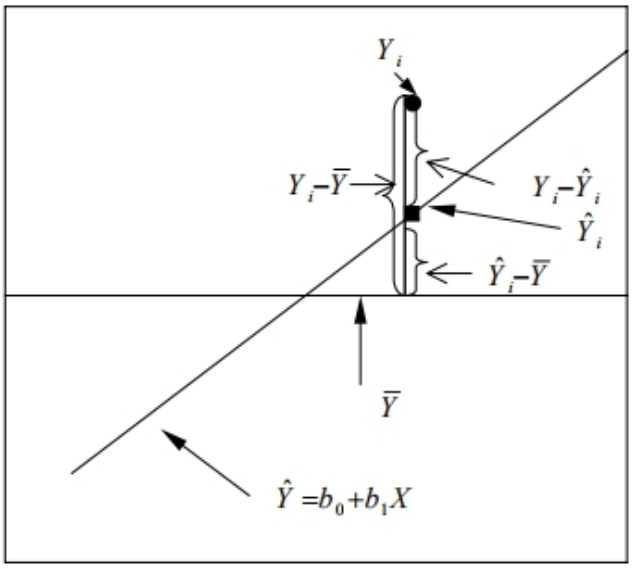
決定係數(coefficient of determination) R^2 是用來解釋線性迴歸模式的適配度(goodness of fit)，R^2 =0 時，代表依變數(Y)與自變數(X_n)沒有線性關係，R^2 ≠0 時， 代表依變數(Y)被自變數(X n)所解釋的比率

R^2 在統計的定義如下：

$$R^2=\frac{SS_{Reg}}{SS_{Total}} = 1 - \frac{SS_{Res}}{SS_{Total}}$$

- SS_{Total} (SST)：總變異量(總平方和)
- SS_{Reg} (SSR)：迴歸可解釋的變易量(迴歸平方和)
- SS_{Res} (SSE)：誤差變異量(殘差平方和)

假設一數據集包括 $y_1,...,y_n$ 共 n 個觀察值，相對應的模型預測值分別為 $\hat{y}_1,...,\hat{y}_n$ ，定義殘差 $e_i = y_i - \hat{y}_i$ ，殘差越小代表愈線性迴歸模型的適配度越高。 $\hat{y}_i$為迴歸模式在 y_i點的預測值，$\bar{y}$ ，所有 y_i值的平均值。



$$Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

兩邊取平方再加總 → $\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + 2\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)$

SST 總平方和　SSR 迴歸平方和　SSE 殘差平方和　根據殘差性質，此交差項為0

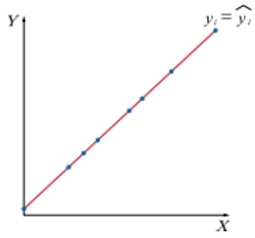
應變數Y的所有變異 = 迴歸模式對Y可以解釋的變異 + 迴歸模式對Y未能解釋的變異

※ 迴歸模式對Y變異的解釋比例 = $\frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST} = R^2$

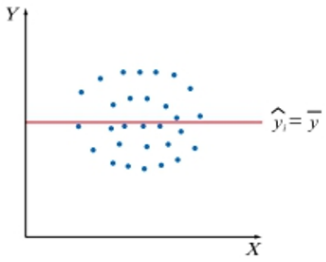
✧ R^2 的定義代表迴歸模式之變異值與所有 yi 變異量之比例，R^2 愈大，代表此迴歸模式能夠解釋全體 yi 變異量的比例愈大。因此 R^2 愈接近 1.0，代表此模式愈有解釋能力。

補充：

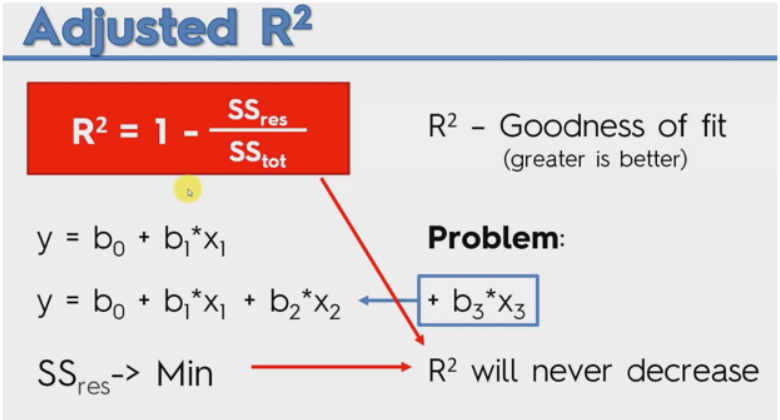
(一) 每個樣本點都落在迴歸直線上，
則每一個資料點的殘差均為0 (SSE=0)
則此時 R^2 =1 (完美配適)
該迴歸直線具有非常強烈的解釋能力
即 Y 的總變異都能被 X 的變異解釋



(二) 若所有樣本迴歸直線上的 $\hat{y}_i = \bar{y}$，
則 SSR=0。
則 此時 $R^2 = 0$ (最差配適)
該迴歸直線完全沒有解釋能力
即 Y 的總變異不能被 X 的變異解釋



二、Adjusted R-Squared Intuition



假設用$X_1, X_2, \ldots, X_p$預測Y，則我們有P個預測變數，需估計P+1個參數。

解釋能力R^2雖然越大越好，但在建模時，還是需考慮精簡原則。

R^2對P是具有單調性。也就是投入越多預測變數，模型解釋能力R^2就會越高，但模型也變的更複雜。

✧ 在迴歸模式中，R^2 會用來說明整個模式的解釋力，但是 R^2 會受到樣本(n)大小與解釋變數(p)的多寡影響而呈現高估現象，樣本愈小或解釋變數越多時，愈容易出現問題(高估)，因此，大多數的學者都採用調整後的 R^2 ，也就是將誤差變異量和依變數(Y)的總變異量都除以自由度 degree of freedom. (df)

$$\text{Adjusted } R^2 = 1 - \frac{\frac{SS_{Res}}{df_{Res}}}{\frac{SS_{Tot}}{df_{Tot}}} = 1 - \frac{\frac{SS_{Res}}{n-p-1}}{\frac{SS_{Tot}}{n-1}}$$

其中 n 為樣本數，p 為解釋變數個數

✧ Adjusted R^2 值越近 1 表示迴歸式越顯著。

✧ Adjusted R^2 可以理解為，給進入模型的輸入變量一個懲罰機制，你加入的輸入變量 X 越多，我的懲罰越大。因此校正 R 平方可以理解為計算真正和 Y 有關的輸入變量 X 可以解釋的 Y 的百分比。它引入了模型的自由度，自由度從統計上來講，是指當以樣本的統計量來估計總體的參數時，樣本中獨立或能自由變化的數據的個數。在這裡為了方便理解，可以認為是代表著引入輸入變量 X 的個數。因此如果加入的輸入變量能解釋的 Y 的百分比無法抗衡自由度的增加的話，你的 Adjusted R^2 不會增加而是會降低。

三、Evaluating Regression Models Performance

- ✓ Model 放入愈多變數，R-square 會愈高，但是 model 會太混亂；而且有些變數加進去卻只有提升一咪咪的 R-square。故建議以 Adj. R-square 做為判定依據。

Report 1.

```
Call:
lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend +
    State, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33504  -4736     90   6672  17338

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.008e+04  6.953e+03   7.204 5.76e-09 ***
R.D.Spend        8.060e-01  4.641e-02  17.369  < 2e-16 ***
Administration  -2.700e-02  5.223e-02  -0.517    0.608
Marketing.Spend  2.698e-02  1.714e-02   1.574    0.123
State2           4.189e+01  3.256e+03   0.013    0.990
State3           2.407e+02  3.339e+03   0.072    0.943
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9439 on 44 degrees of freedom
Multiple R-squared:  0.9508,    Adjusted R-squared:  0.9452
F-statistic: 169.9 on 5 and 44 DF,  p-value: < 2.2e-16
```

變數：5　　R-squared：0.9508　　Adj R-squared：0.9452

Report 2.

```
Call:
lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend,
    data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33534  -4795     63   6606  17275

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.012e+04  6.572e+03   7.626 1.06e-09 ***
R.D.Spend        8.057e-01  4.515e-02  17.846  < 2e-16 ***
Administration  -2.682e-02  5.103e-02  -0.526    0.602
Marketing.Spend  2.723e-02  1.645e-02   1.655    0.105
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9232 on 46 degrees of freedom
Multiple R-squared:  0.9507,    Adjusted R-squared:  0.9475
F-statistic:   296 on 3 and 46 DF,  p-value: < 2.2e-16
```

變數：3　　R-squared：0.9507　　Adj R-squared：0.9475

Report 3.

```
Call:
lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33645  -4632   -414   6484  17097

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.698e+04  2.690e+03  17.464   <2e-16 ***
R.D.Spend       7.966e-01  4.135e-02  19.266   <2e-16 ***
Marketing.Spend 2.991e-02  1.552e-02   1.927     0.06 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9161 on 47 degrees of freedom
Multiple R-squared:  0.9505,    Adjusted R-squared:  0.9483
F-statistic: 450.8 on 2 and 47 DF,  p-value: < 2.2e-16
```

變數：2　　R-squared：0.9505　　Adj R-squared：0.9483

Report 4.

```
Call:
lm(formula = Profit ~ R.D.Spend, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-34351  -4626    -375   6249  17188

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.903e+04  2.538e+03   19.32   <2e-16 ***
R.D.Spend   8.543e-01  2.931e-02   29.15   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9416 on 48 degrees of freedom
Multiple R-squared:  0.9465,    Adjusted R-squared:  0.9454
F-statistic: 849.8 on 1 and 48 DF,  p-value: < 2.2e-16
```

變數：1　　R-squared：0.9465　　Adj R-squared：0.9454

- **變數越多，R-squared 越大**
- **從 Adj R-squared 來選擇最配適的 MODEL，因此選擇 Report 3**

四、Interpreting Linear Regression Coefficients

選擇第三個模型(Adjusted R^2 最大)為配適最佳模型

```
Call:
lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33645  -4632   -414   6484  17097

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.698e+04  2.690e+03  17.464   <2e-16 ***
R.D.Spend       7.966e-01  4.135e-02  19.266   <2e-16 ***
Marketing.Spend 2.991e-02  1.552e-02   1.927     0.06 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9161 on 47 degrees of freedom
Multiple R-squared:  0.9505,    Adjusted R-squared:  0.9483
F-statistic: 450.8 on 2 and 47 DF,  p-value: < 2.2e-16
```

接下來開始觀察 Coefficient 估計量

1. 估計量>0　代表此變數與你的獨立變數相關，並 collate 相依變數

   意思就是改變獨立變數，相依變數會朝同個方向改變

2. 估計值越大，影響越大

```
lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33645  -4632   -414   6484  17097

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.698e+04  2.690e+03  17.464   <2e-16 ***
R.D.Spend      7.966e-01  4.135e-02  19.266   <2e-16 ***
Marketing.Spend 2.991e-02  1.552e-02   1.927     0.06 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9161 on 47 degrees of freedom
Multiple R-squared:  0.9505,    Adjusted R-squared:  0.9483
F-statistic: 450.8 on 2 and 47 DF,  p-value: < 2.2e-16
```

說明：

1. 每增加一單位 R.D Spend，預期 Profit 會上升 7.966e-01 單位；
2. 每增加一單位 Marketing Spend，預期 Profit 會上升 2.991e-02 單位。
3. 有些公司會認為投入過多經費在 Marketing，卻只有上升 2.991e-02 的 Profit，很不划算。但這屬於公司決策，我們要做的只是建立 model。