機器學習
Polynomial Regression

授課老師：林彥廷

多項式回歸
Polynomial Regression

- 研究一個因變數與一個或多個自變量間多項式的回歸分析方法，稱為多項式回歸（Polynomial Regression）。如果自變量只有一個時，稱為一元多項式回歸；如果自變量有多個時，稱為多元多項式回歸。在一元回歸分析中，如果因變數y與自變數x的關係為非線性的，但是又找不到適當的函數曲線來擬合，則可以採用一元多項式回歸。

回歸 Regressions

Simple Linear Regression

$$y = b_0 + b_1 x_1$$

回歸 Regressions

Simple Linear Regression

$$y = b_0 + b_1 x_1$$

Multiple Linear Regression

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

回歸 Regressions

Simple Linear Regression

$$y = b_0 + b_1 x_1$$

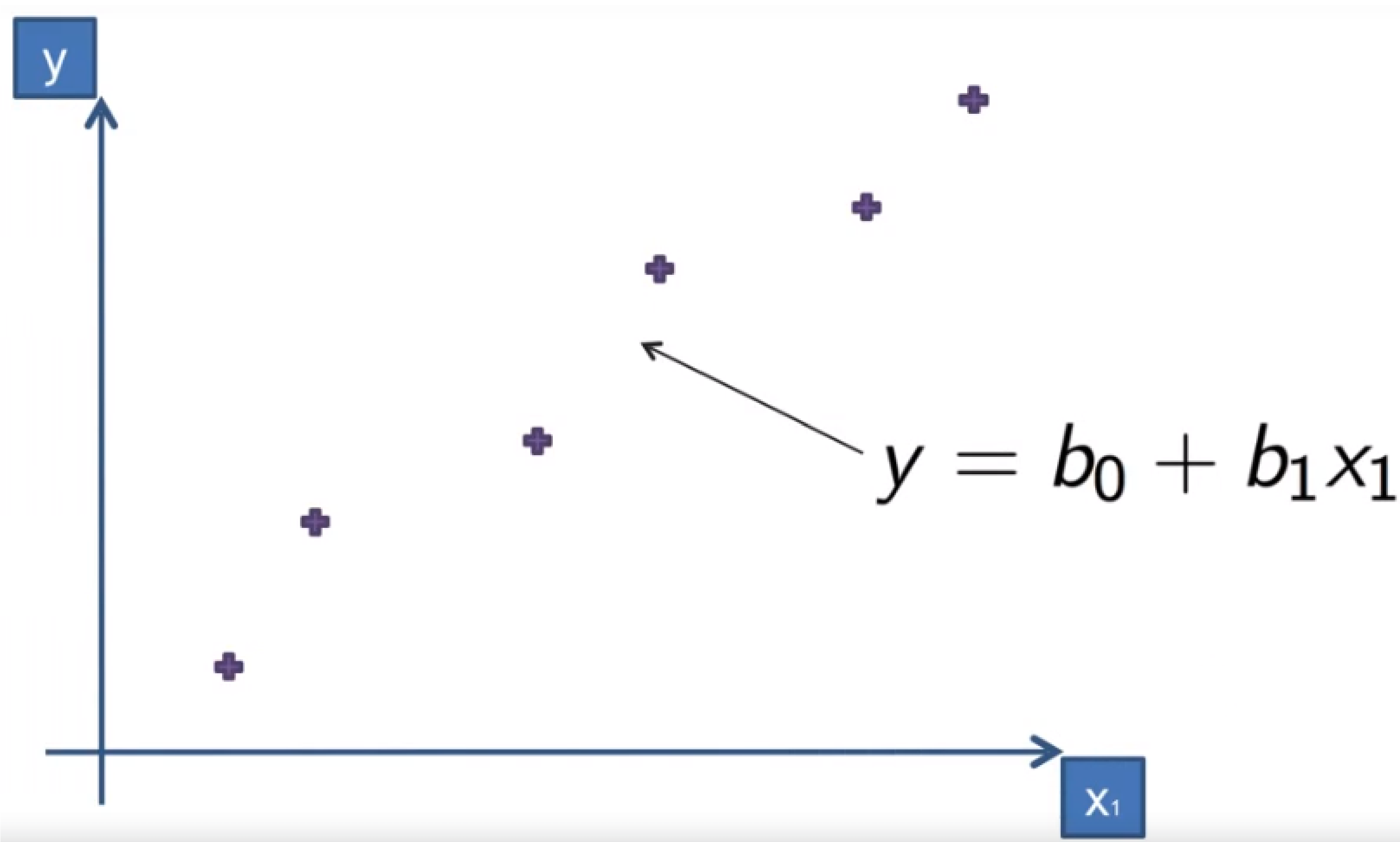
Multiple Linear Regression

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

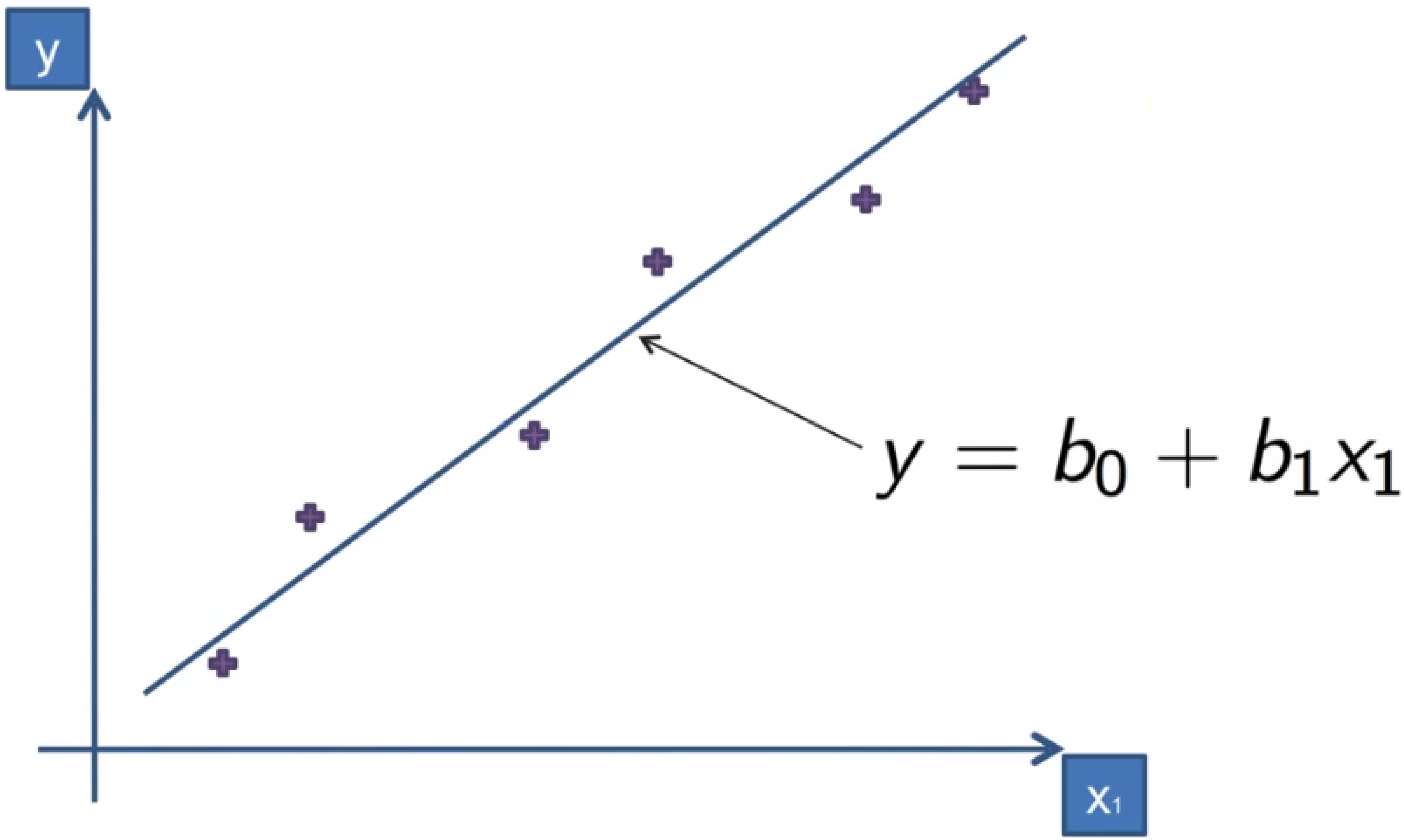
Polynomial Linear Regression

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$$

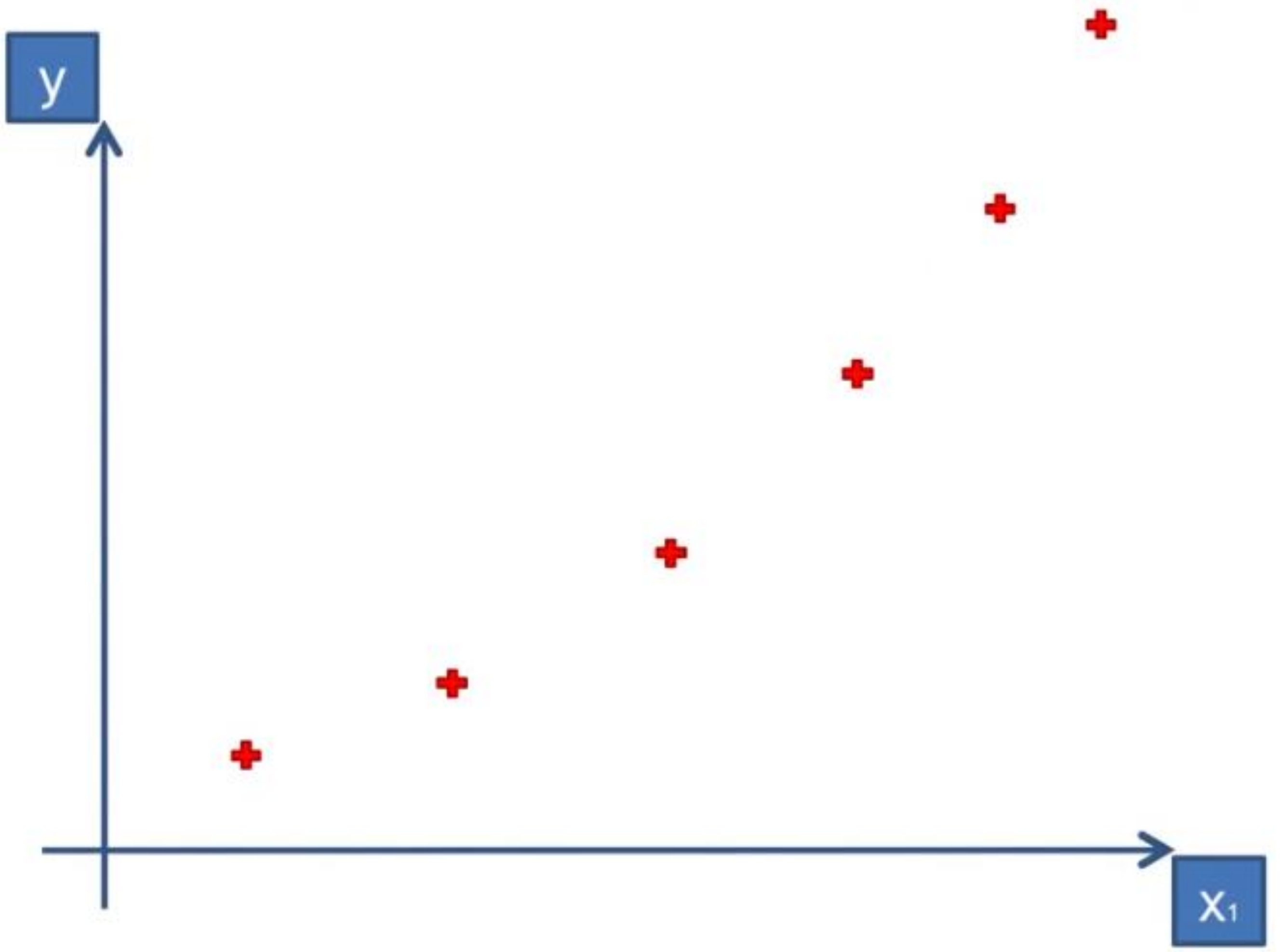
Simple Linear Regression



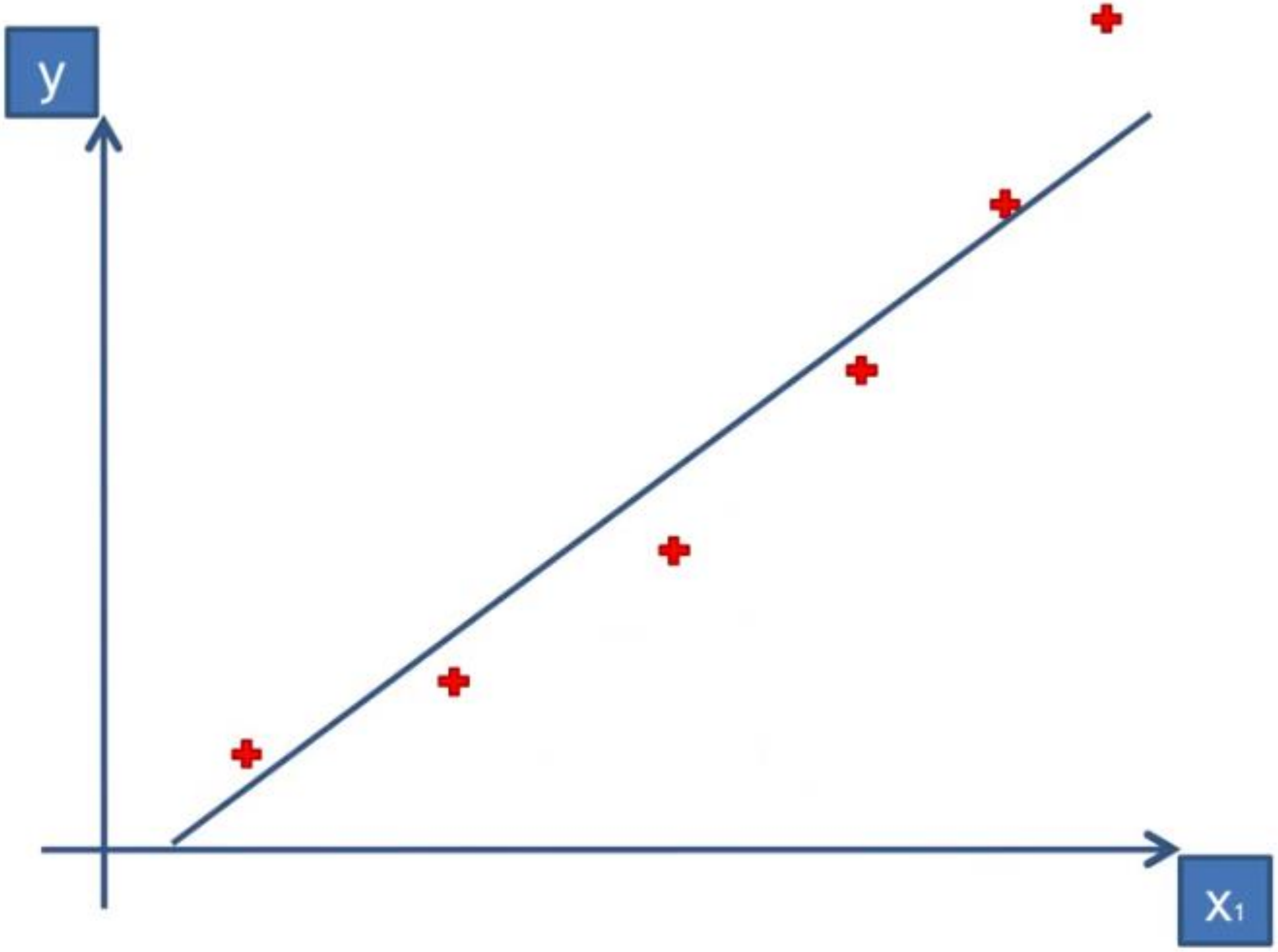
Simple Linear Regression



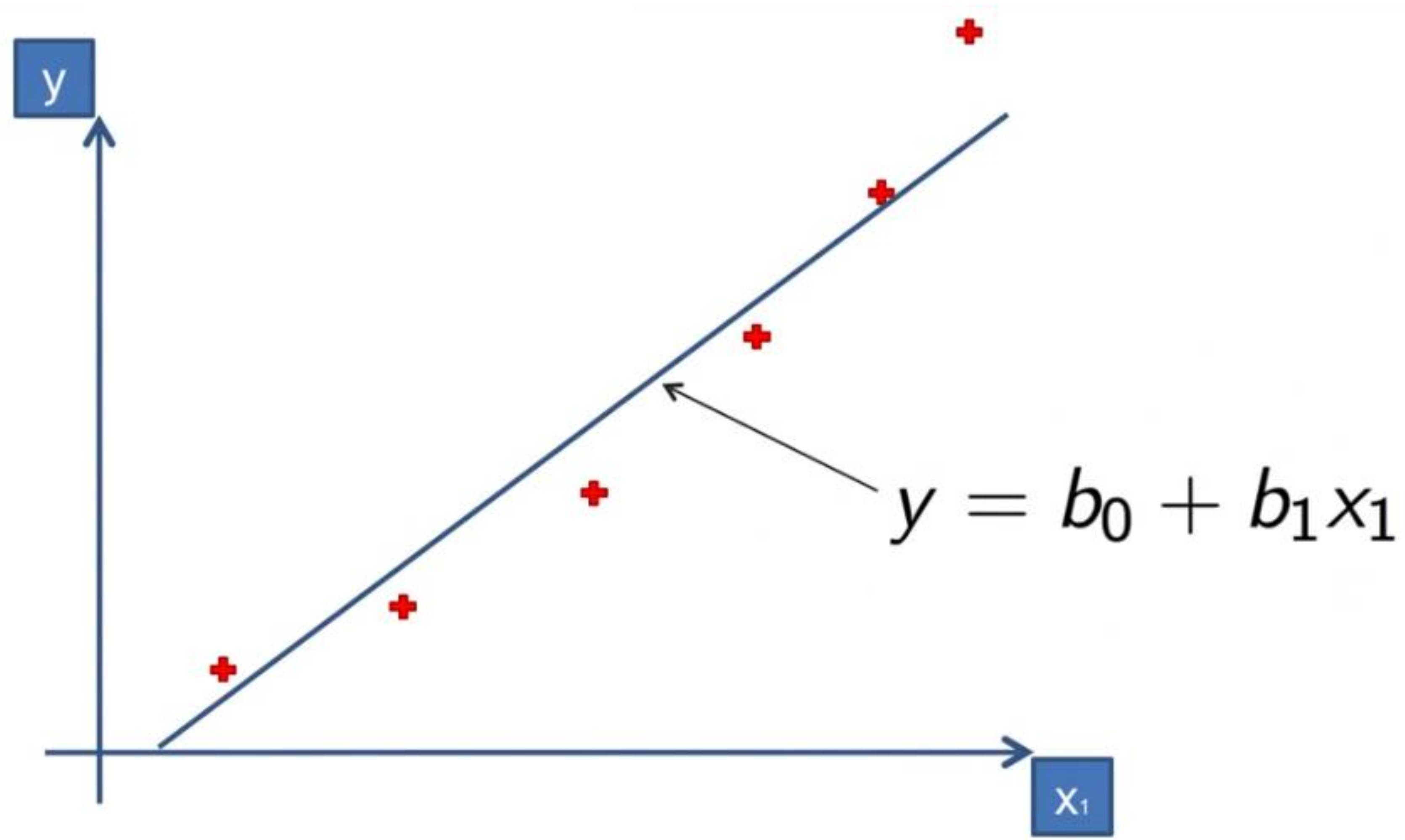
Simple Linear Regression



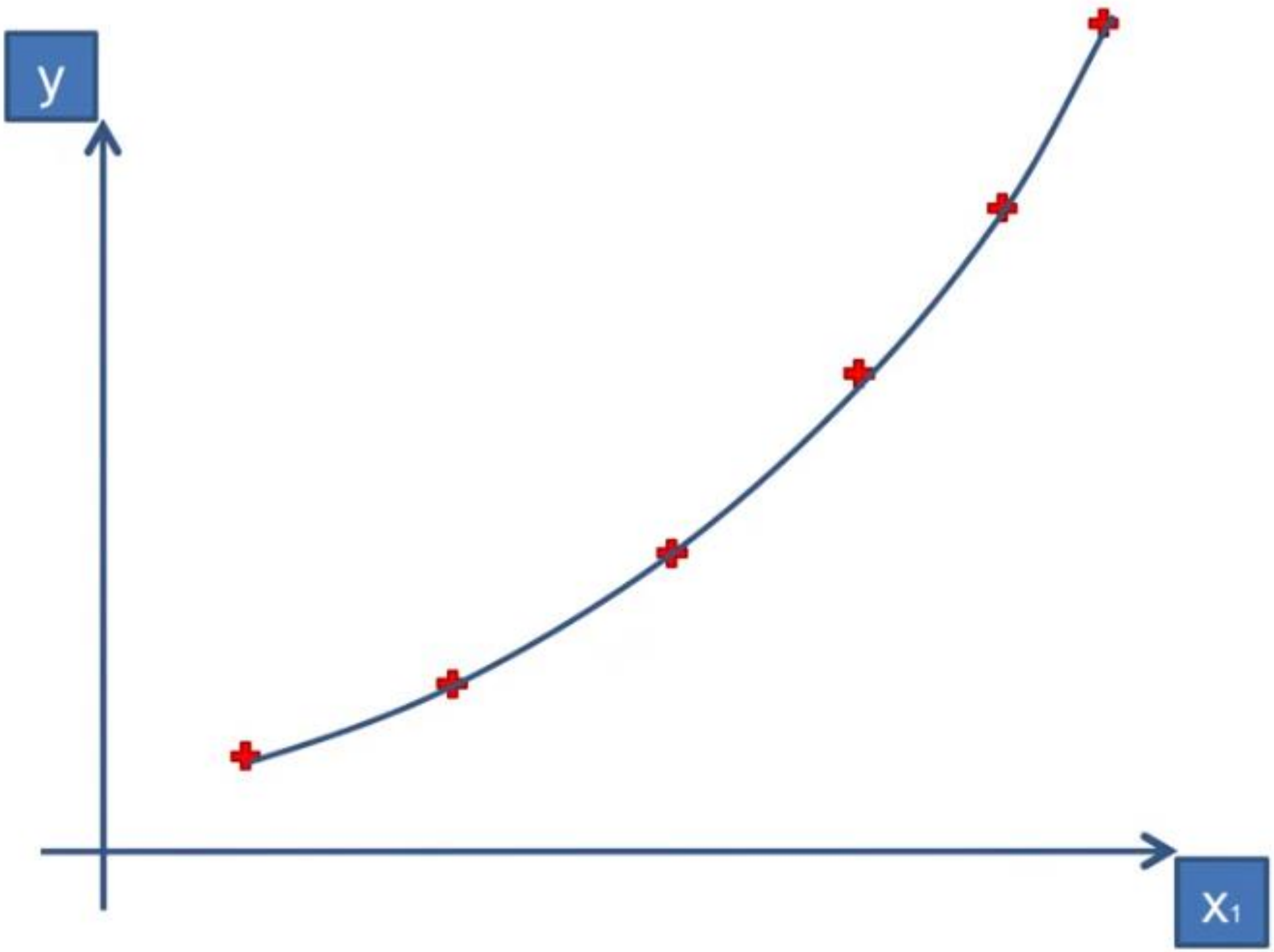
Simple Linear Regression



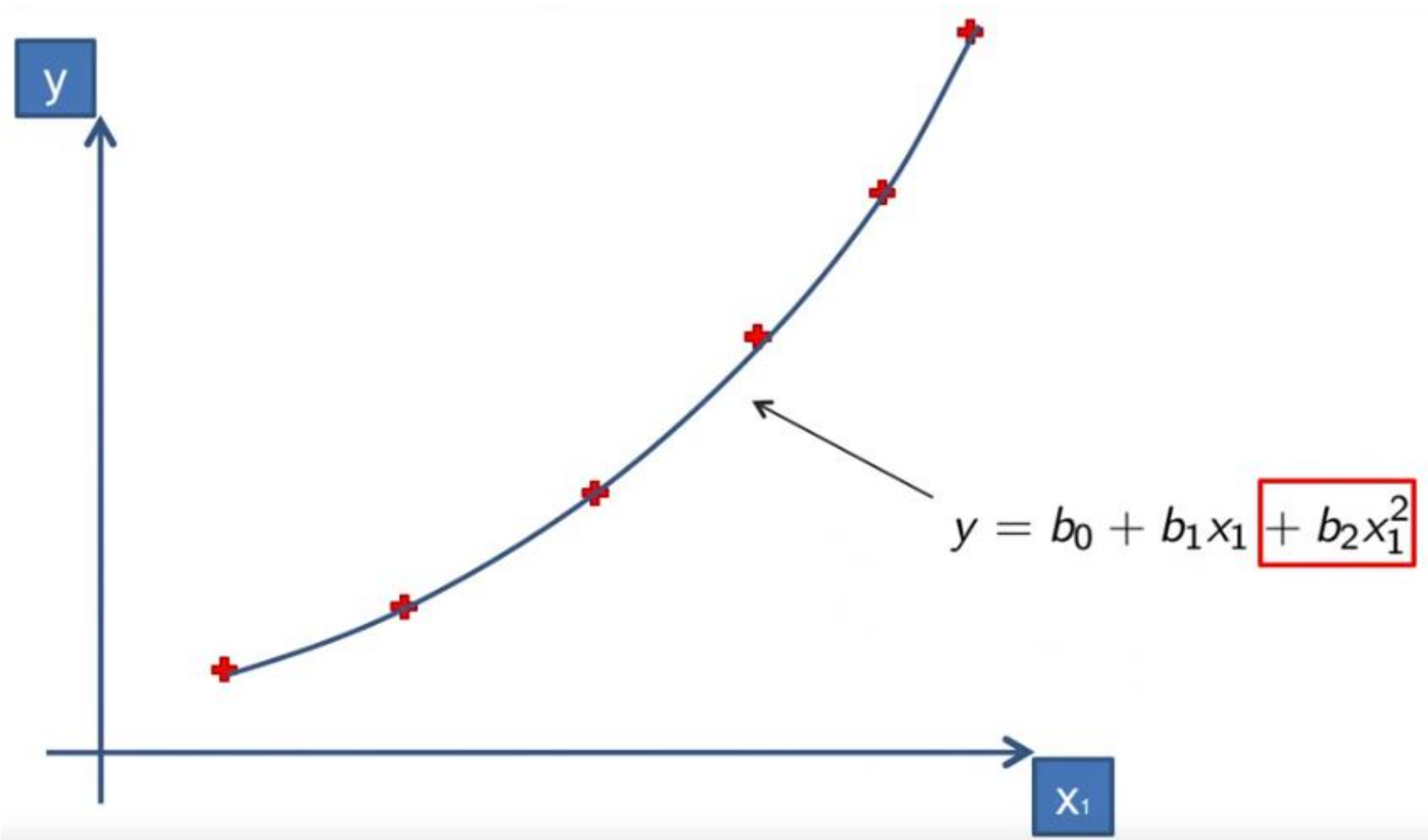
Simple Linear Regression



Polynomial Regression



Polynomial Regression



Polynomial Regression

One Question: Why "Linear"?

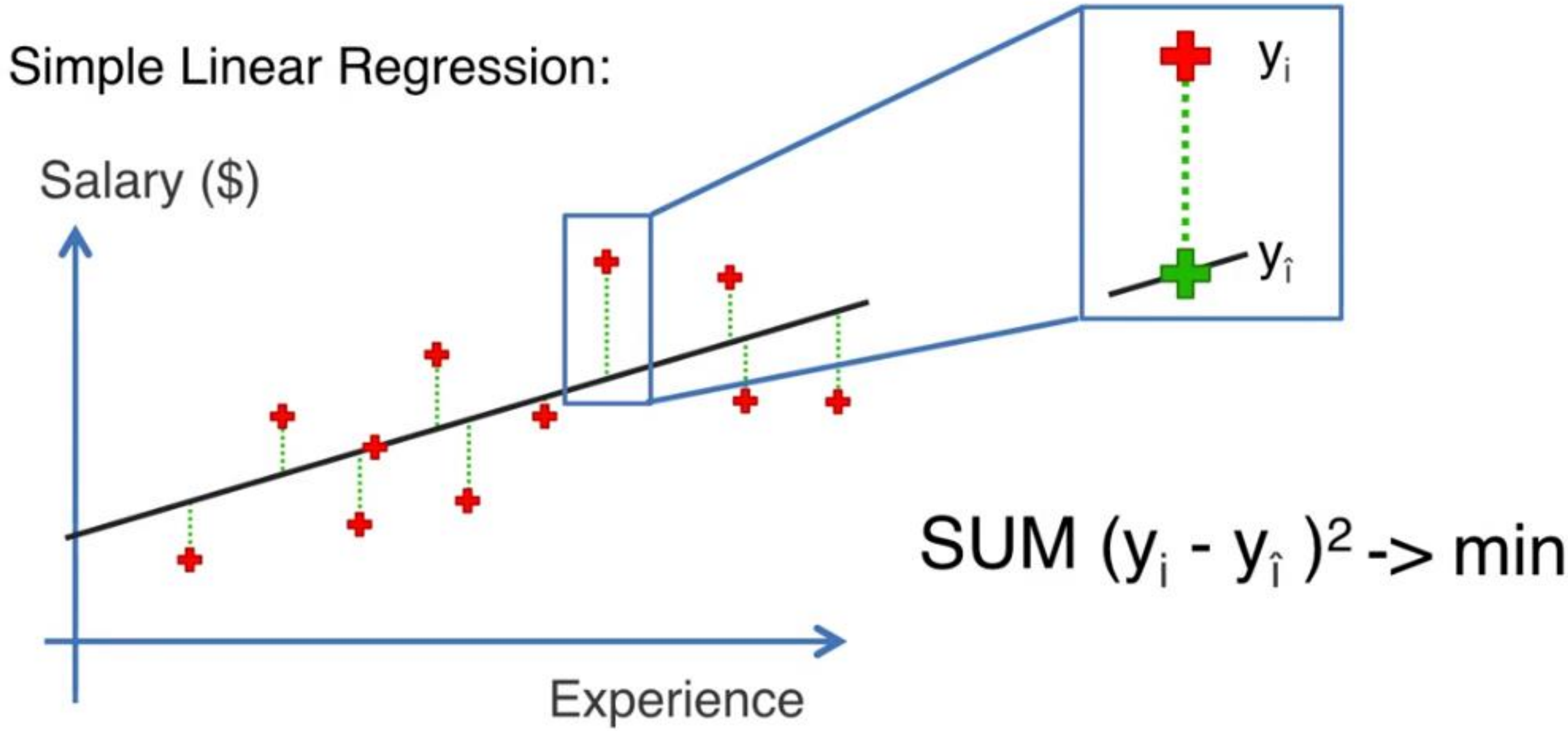
Polynomial Regression

Polynomial Linear Regression

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$$

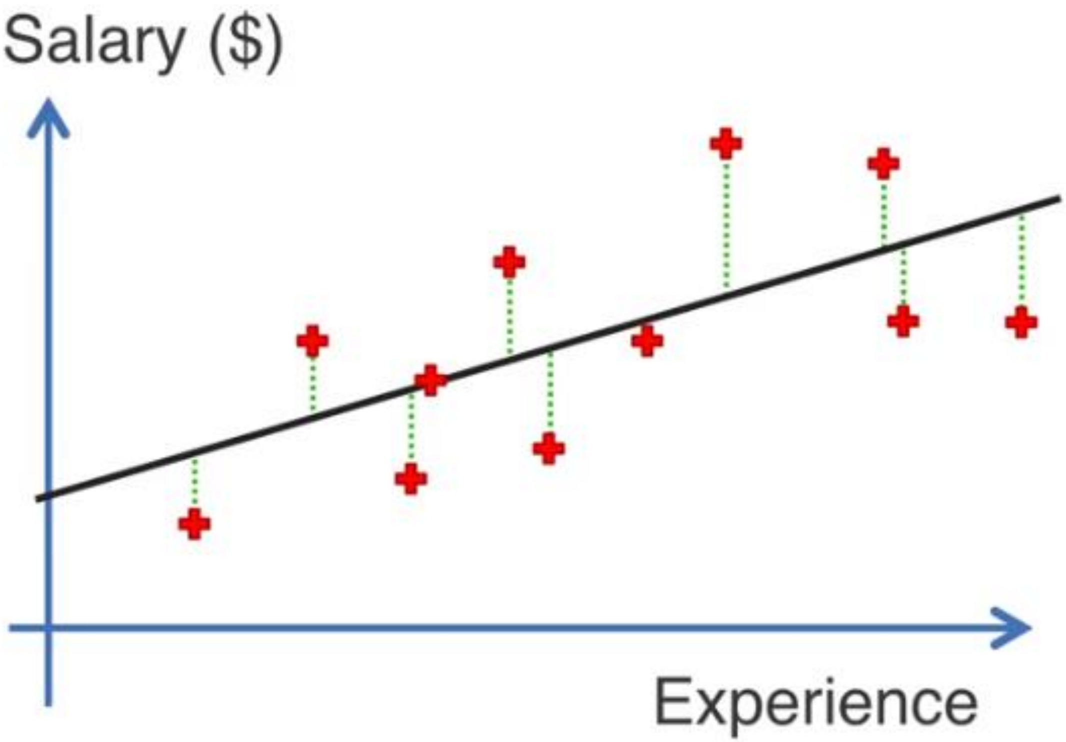
R Squared Intuition

R Squared


Simple Linear Regression:
Salary ($)
Experience
y_i
$\hat{y}_i$
SUM $(y_i - \hat{y}_i)^2$ -> min

R Squared

Simple Linear Regression:

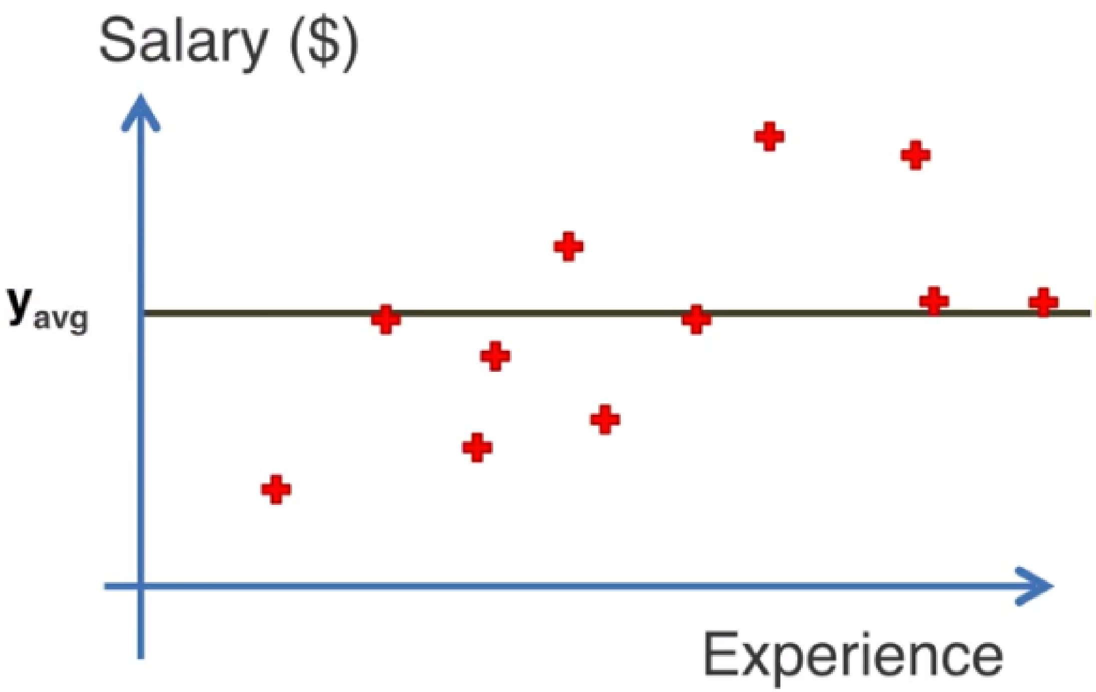


殘差平方和

$$SS_{res} = SUM\,(y_i - y_{\hat{i}})^2$$

R Squared

Simple Linear Regression:



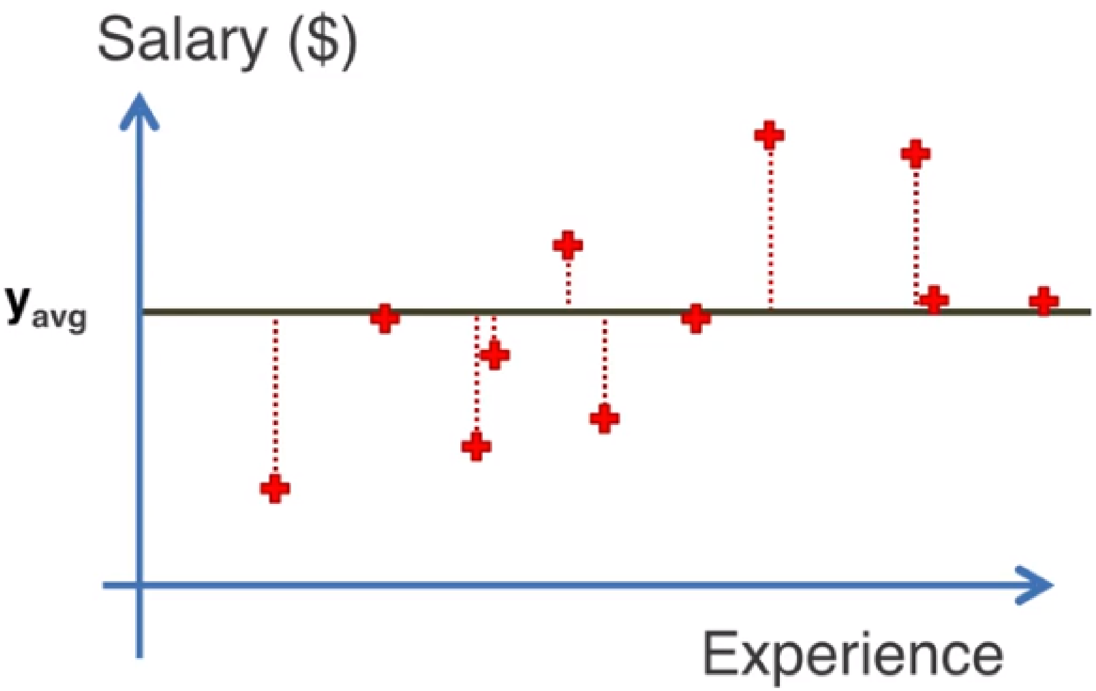
殘差平方和

$SS_{res} = SUM\ (y_i - y_{\hat{i}})^2$

R Squared

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Simple Linear Regression:



殘差平方和

SS_{res} =SUM $(y_i - y_{\hat{i}})^2$

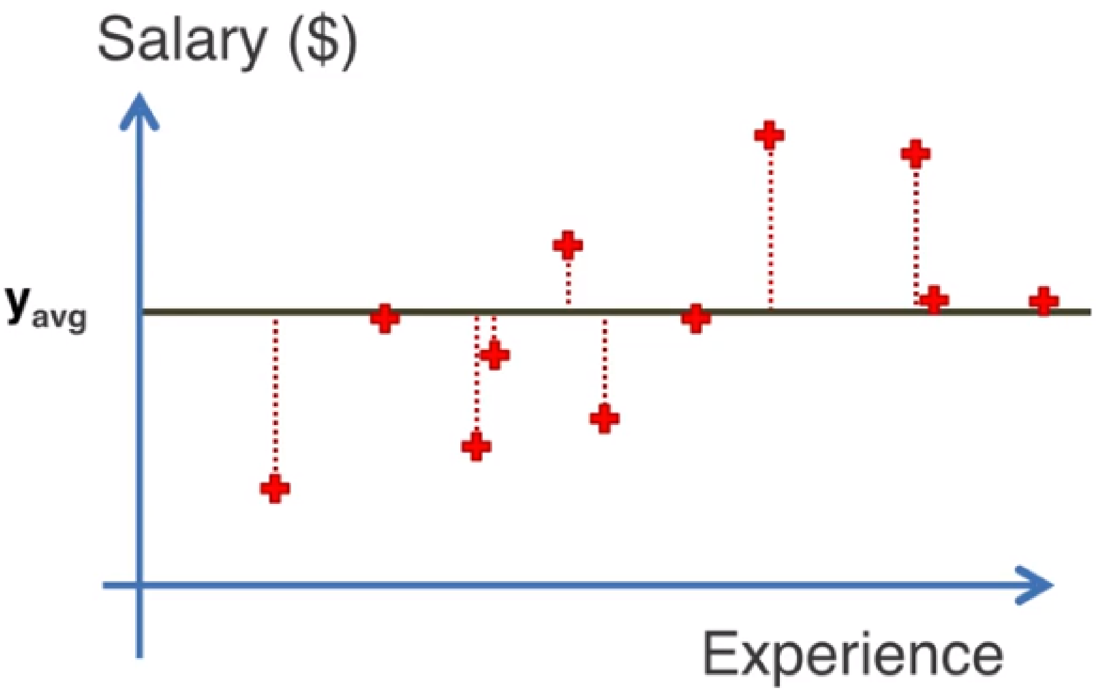
總平方和

SS_{tot} =SUM $(y_i - y_{avg})^2$

R Squared

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Simple Linear Regression:



殘差平方和

SS_{res} =SUM $(y_i - y_{\hat{i}})^2$

總平方和

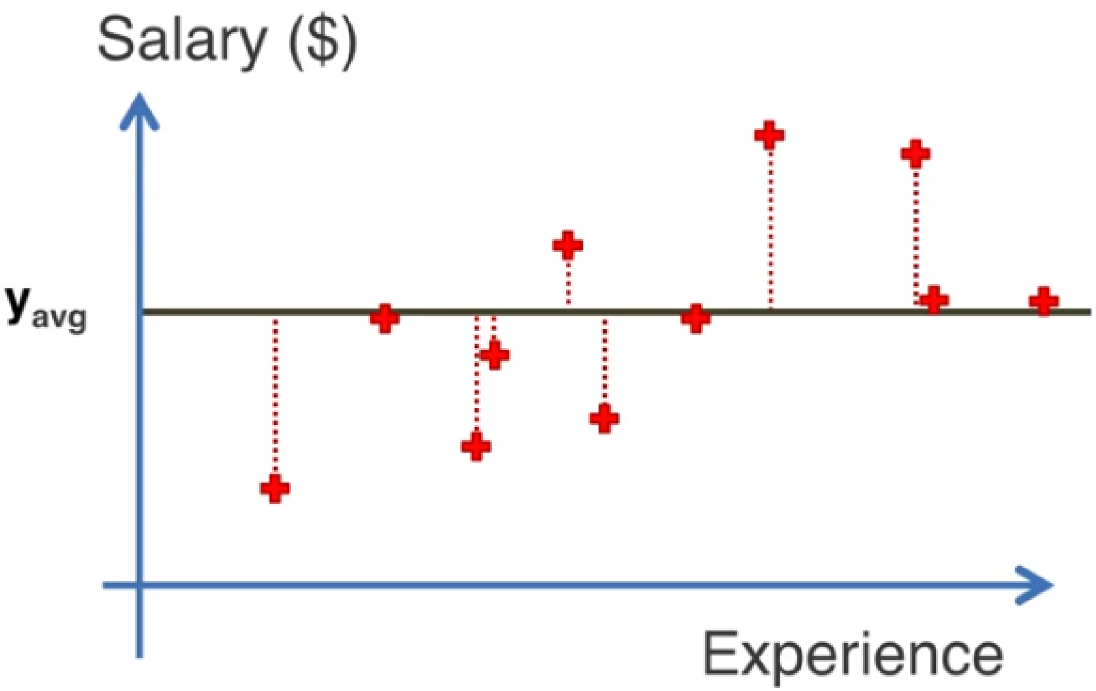
SS_{tot} =SUM $(y_i - y_{avg})^2$

Adjusted R^2 Intuition

Adjusted R Squared

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Simple Linear Regression:



殘差平方和

$SS_{res} = SUM\ (y_i - y_{\hat{i}})^2$

總平方和

$SS_{tot} = SUM\ (y_i - y_{avg})^2$

Adjusted R Squared

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

R^2 – Goodness of fit
(greater is better)

$y = b_0 + b_1{*}x_1$

$y = b_0 + b_1{*}x_1 + b_2{*}x_2$

SS_{res} -> Min

Adjusted R Squared

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

R^2 – Goodness of fit
(greater is better)

$y = b_0 + b_1 * x_1$

Problem:

$y = b_0 + b_1 * x_1 + b_2 * x_2 \leftarrow \boxed{+ b_3 * x_3}$

SS_{res} -> Min

Adjusted R Squared

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

R^2 – Goodness of fit

(greater is better)

$y = b_0 + b_1{}^*x_1$

Problem:

$y = b_0 + b_1{}^*x_1 + b_2{}^*x_2 \leftarrow \boxed{+ b_3{}^*x_3}$

SS_{res} -> Min $\longrightarrow$ R^2 will never decrease

Adjusted R Squared

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

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

p - number of regressors 自變量個數

n – sample size 資料個數

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

Residuals:
   Min     1Q Median     3Q    Max
-33128  -4865      5   6098  18065

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.965e+04  7.637e+03   6.501 1.94e-07 ***
R.D.Spend       7.986e-01  5.604e-02  14.251 6.70e-16 ***
Administration -2.942e-02  5.828e-02  -0.505    0.617
Marketing.Spend 3.268e-02  2.127e-02   1.537    0.134
State2          1.213e+02  3.751e+03   0.032    0.974
State3          2.376e+02  4.127e+03   0.058    0.954
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9908 on 34 degrees of freedom
Multiple R-squared:  0.9499,	Adjusted R-squared:  0.9425
F-statistic:   129 on 5 and 34 DF,  p-value: < 2.2e-16
```

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

Residuals:
   Min     1Q Median     3Q    Max
-33117  -4858    -36   6020  17957

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.970e+04  7.120e+03   6.980 3.48e-08 ***
R.D.Spend        7.983e-01  5.356e-02  14.905  < 2e-16 ***
Administration  -2.895e-02  5.603e-02  -0.517    0.609
Marketing.Spend  3.283e-02  1.987e-02   1.652    0.107
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9629 on 36 degrees of freedom
Multiple R-squared:  0.9499,	Adjusted R-squared:  0.9457
F-statistic: 227.6 on 3 and 36 DF,  p-value: < 2.2e-16
```

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

Residuals:
   Min     1Q Median     3Q    Max
-33294  -4763   -354   6351  17693

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.638e+04  3.019e+03  15.364   <2e-16 ***
R.D.Spend       7.879e-01  4.916e-02  16.026   <2e-16 ***
Marketing.Spend 3.538e-02  1.905e-02   1.857   0.0713 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9533 on 37 degrees of freedom
Multiple R-squared:  0.9495,	Adjusted R-squared:  0.9468
F-statistic: 348.1 on 2 and 37 DF,  p-value: < 2.2e-16
```

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

Residuals:
   Min     1Q Median     3Q    Max
-34334  -4894   -340   6752  17147

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.902e+04  2.748e+03   17.84   <2e-16 ***
R.D.Spend   8.563e-01  3.357e-02   25.51   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9836 on 38 degrees of freedom
Multiple R-squared:  0.9448,	Adjusted R-squared:  0.9434
F-statistic: 650.8 on 1 and 38 DF,  p-value: < 2.2e-16
```

THE END

ytlin@mail.nptu.edu.tw