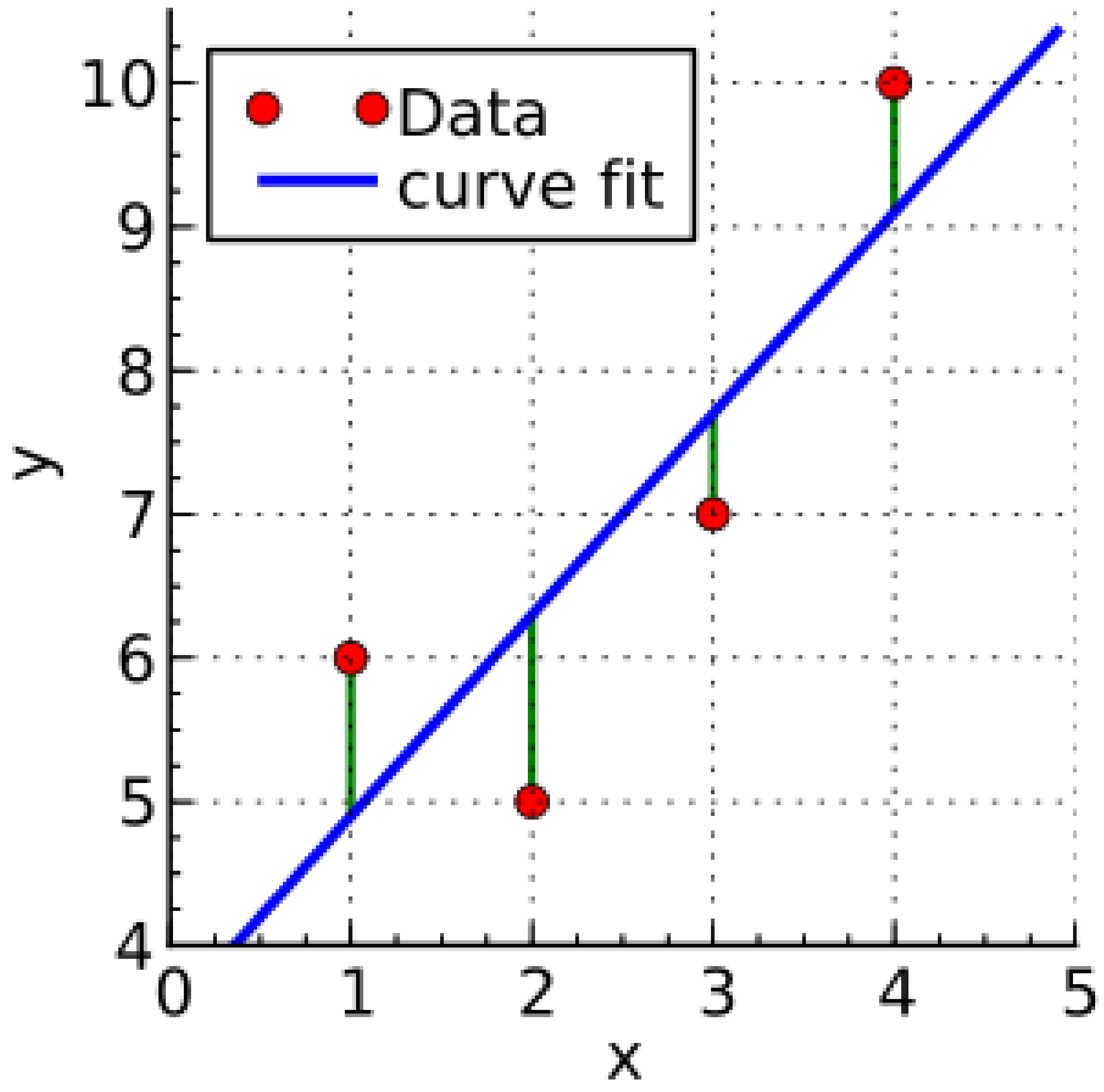
Linear Regression

姚博瀚

- **Simple Linear Regression**
- **Cost Function & Sum of squared residuals**
- **Ordinary least squares method**
- **Linear Regression in R**

Simple Linear Regression

Simple Linear Regression



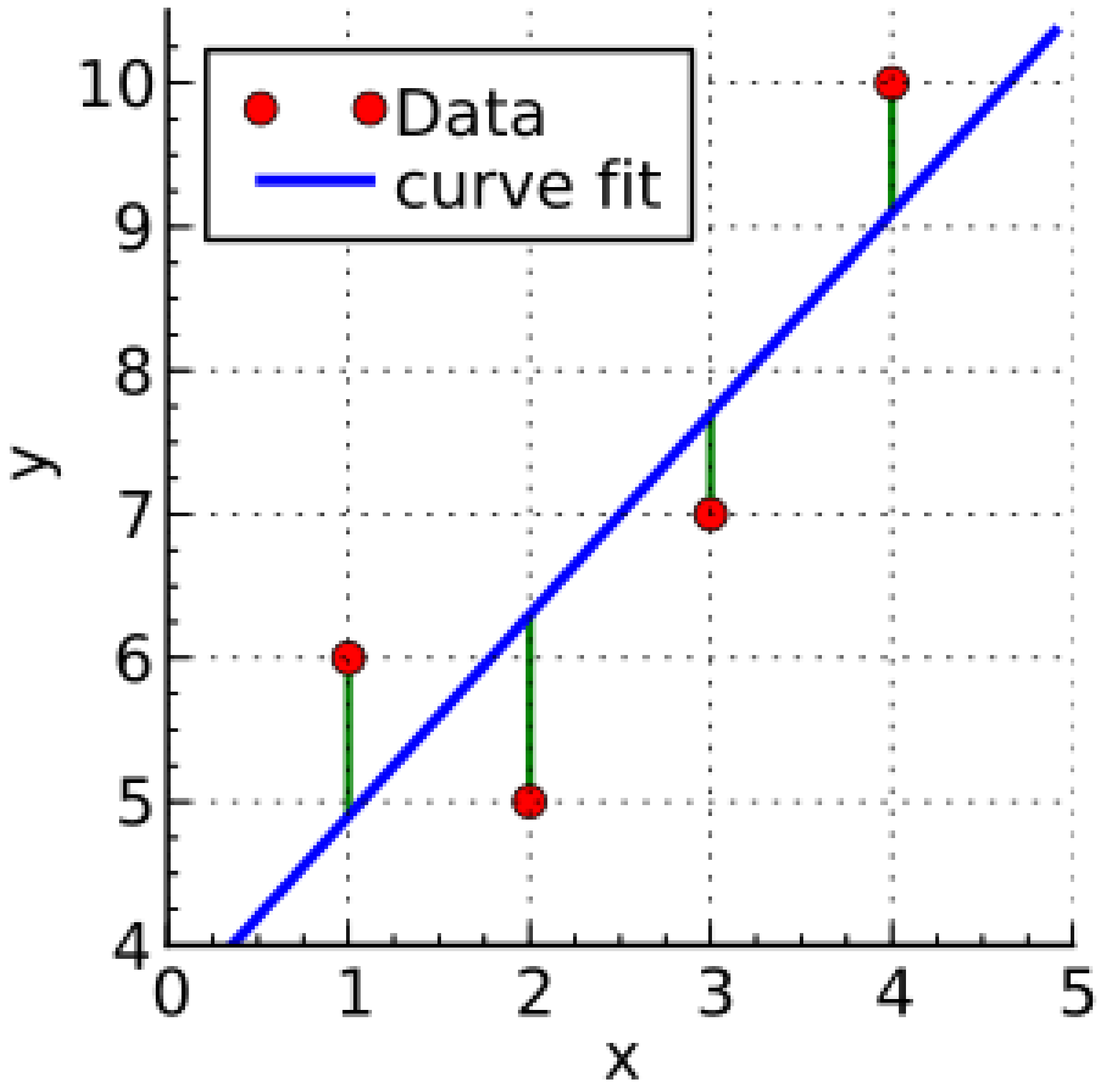
Given four (x, y) data points:
(1,6), (2,5), (3,7), (4,10)

Formula:

$$y = \beta_1 + \beta_2 x$$

Cost Function & Sum of squared residuals

Cost Function & Sum of squared residuals

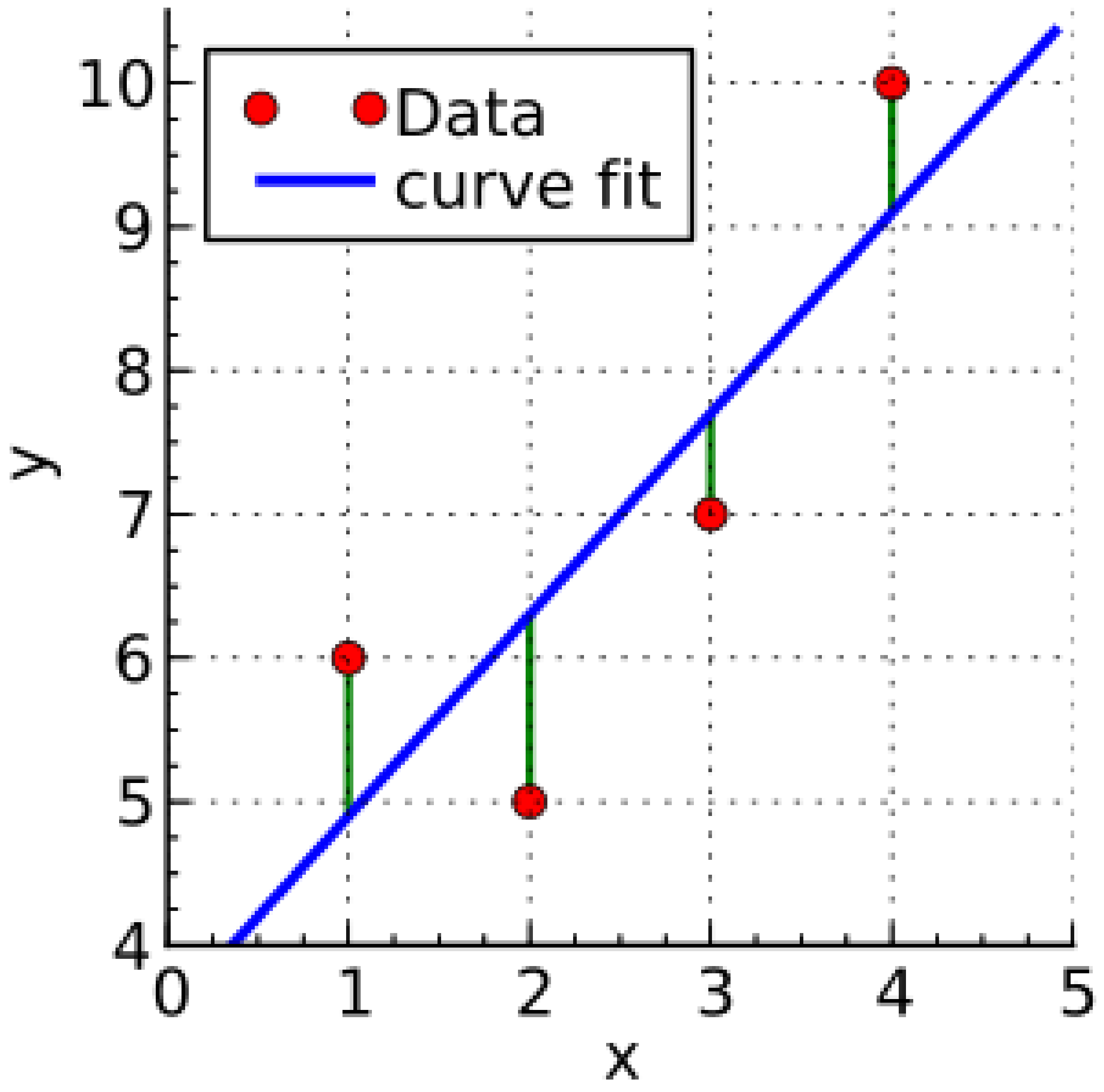


$$CF\left(\beta_1, \beta_2\right) = \frac{1}{2m} \sum_{i=1}^{n} \left(y\left(x_i\right) - y_i\right)^2$$

$$S = \sum_{i=1}^{n} \left(y\left(x_i\right) - y_i\right)^2 = \sum_{i=1}^{n} r_i^2$$

Ordinary least squares method

Ordinary least squares method

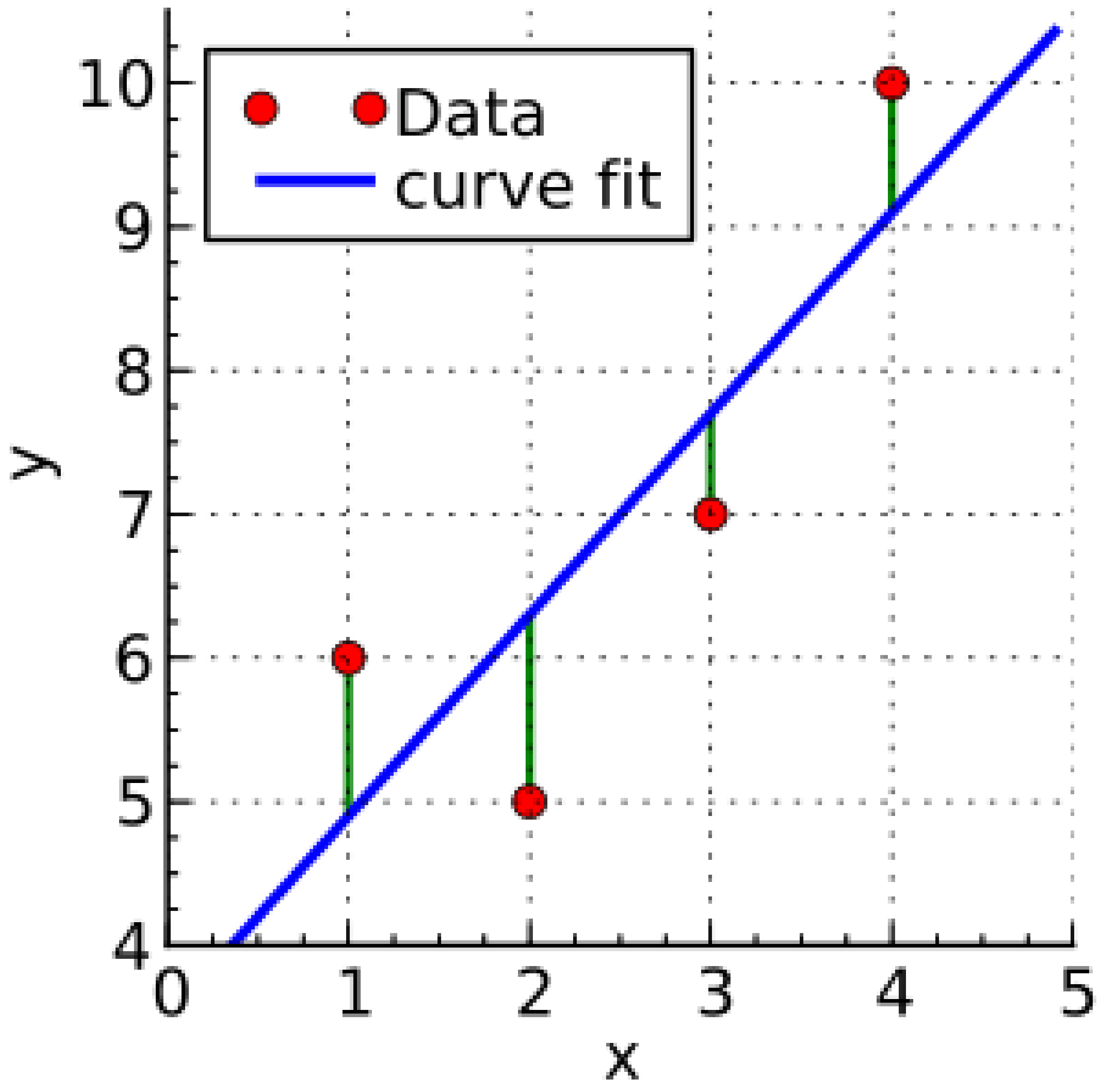


Goal:

Find $\min\left(S\right)$

$$S = \sum_{i=1}^{n} \left(y\left(x_i\right) - y_i\right)^2 = \sum_{i=1}^{n} r_i^2$$

Ordinary least squares method



$$\begin{aligned}
\beta_1 + 1\beta_2 &= 6 \\
\beta_1 + 2\beta_2 &= 5 \\
\beta_1 + 3\beta_2 &= 7 \\
\beta_1 + 4\beta_2 &= 10
\end{aligned}$$

$$\begin{aligned}
\beta_1 + 1\beta_2 + r_1 &= 6 \\
\beta_1 + 2\beta_2 + r_2 &= 5 \\
\beta_1 + 3\beta_2 + r_3 &= 7 \\
\beta_1 + 4\beta_2 + r_4 &= 10
\end{aligned}$$

$$\begin{aligned}
r_1 &= 6 - (\beta_1 + 1\beta_2) \\
r_2 &= 5 - (\beta_1 + 2\beta_2) \\
r_3 &= 7 - (\beta_1 + 3\beta_2) \\
r_4 &= 10 - (\beta_1 + 4\beta_2)
\end{aligned}$$

Ordinary least squares method

$$S = \sum_{i=1}^{n} \left(y\left(x_i\right) - y_i\right)^2 = \sum_{i=1}^{n} r_i^2$$

$$S\left(\beta_1, \beta_2\right) = r_1^2 + r_2^2 + r_3^2 + r_4^2$$
$$= \left[6 - \left(\beta_1 + 1\beta_2\right)\right]^2 + \left[5 - \left(\beta_1 + 2\beta_2\right)\right]^2 + \left[7 - \left(\beta_1 + 3\beta_2\right)\right]^2 + \left[10 - \left(\beta_1 + 4\beta_2\right)\right]^2$$
$$= 4\beta_1^2 + 30\beta_2^2 + 20\beta_1\beta_2 - 56\beta_1 - 154\beta_2 + 210$$

$$0 = \frac{\partial S}{\partial \beta_1} = 8\beta_1 + 20\beta_2 - 56,$$

$$0 = \frac{\partial S}{\partial \beta_2} = 20\beta_1 + 60\beta_2 - 154.$$

$$\begin{cases} \beta_1 = 3.5 \\ \beta_2 = 1.4 \end{cases}$$

Supplement: Second-partials test

$$A = \frac{\partial^2}{\partial \beta_1^2} S\left(\beta_1, \beta_2\right) = \frac{\partial}{\partial \beta_1}\left(\frac{\partial}{\partial \beta_1} S\left(\beta_1, \beta_2\right)\right) = \frac{\partial}{\partial \beta_1}\left(8\beta_1 + 20\beta_2 - 56\right) = 8$$

$$B = \frac{\partial^2}{\partial \beta_1 \partial \beta_2} S\left(\beta_1, \beta_2\right) = \frac{\partial}{\partial \beta_1}\left(\frac{\partial}{\partial \beta_2} S\left(\beta_1, \beta_2\right)\right) = \frac{\partial}{\partial \beta_1}\left(20\beta_1 + 60\beta_2 + 154\right) = 20$$

$$C = \frac{\partial^2}{\partial \beta_2^2} S\left(\beta_1, \beta_2\right) = \frac{\partial}{\partial \beta_2}\left(20\beta_1 + 60\beta_2 + 154\right) = 20$$

$$D = AC - B^2 = 8 \times 20 - 20^2 = 160 - 400 = -240 < 0$$

Since D < 0, (β1, β2) is a saddle point.

In matrix notation:

$$\mathbf{y} = \mathbf{X}\beta \qquad \mathbf{y} = \begin{bmatrix} 6 \\ 5 \\ 7 \\ 10 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\beta \quad \Rightarrow \quad \mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top \mathbf{X}\beta \quad \Rightarrow \quad \beta = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y} = \begin{bmatrix} 3.5 \\ 1.4 \end{bmatrix}$$

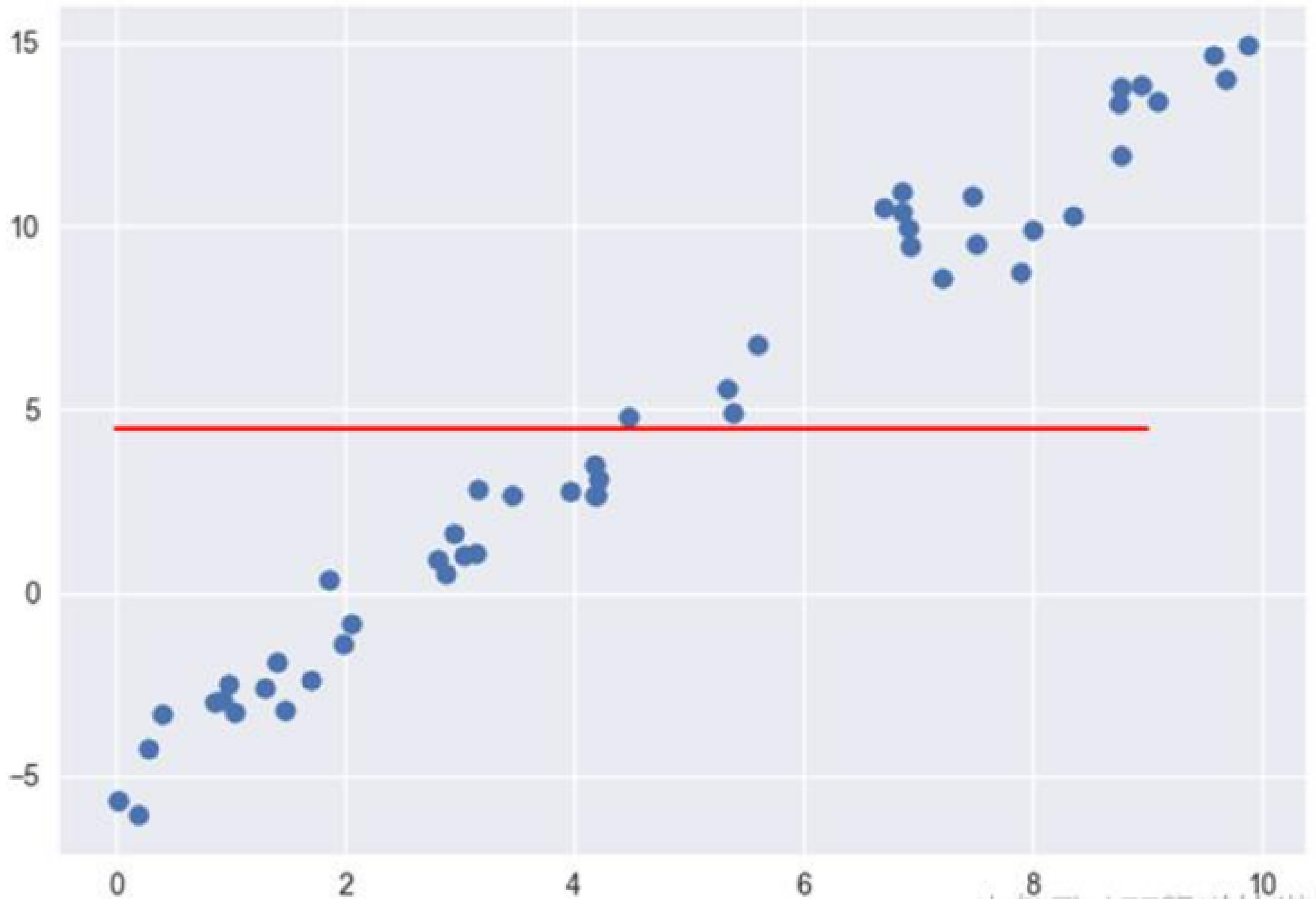
R-square

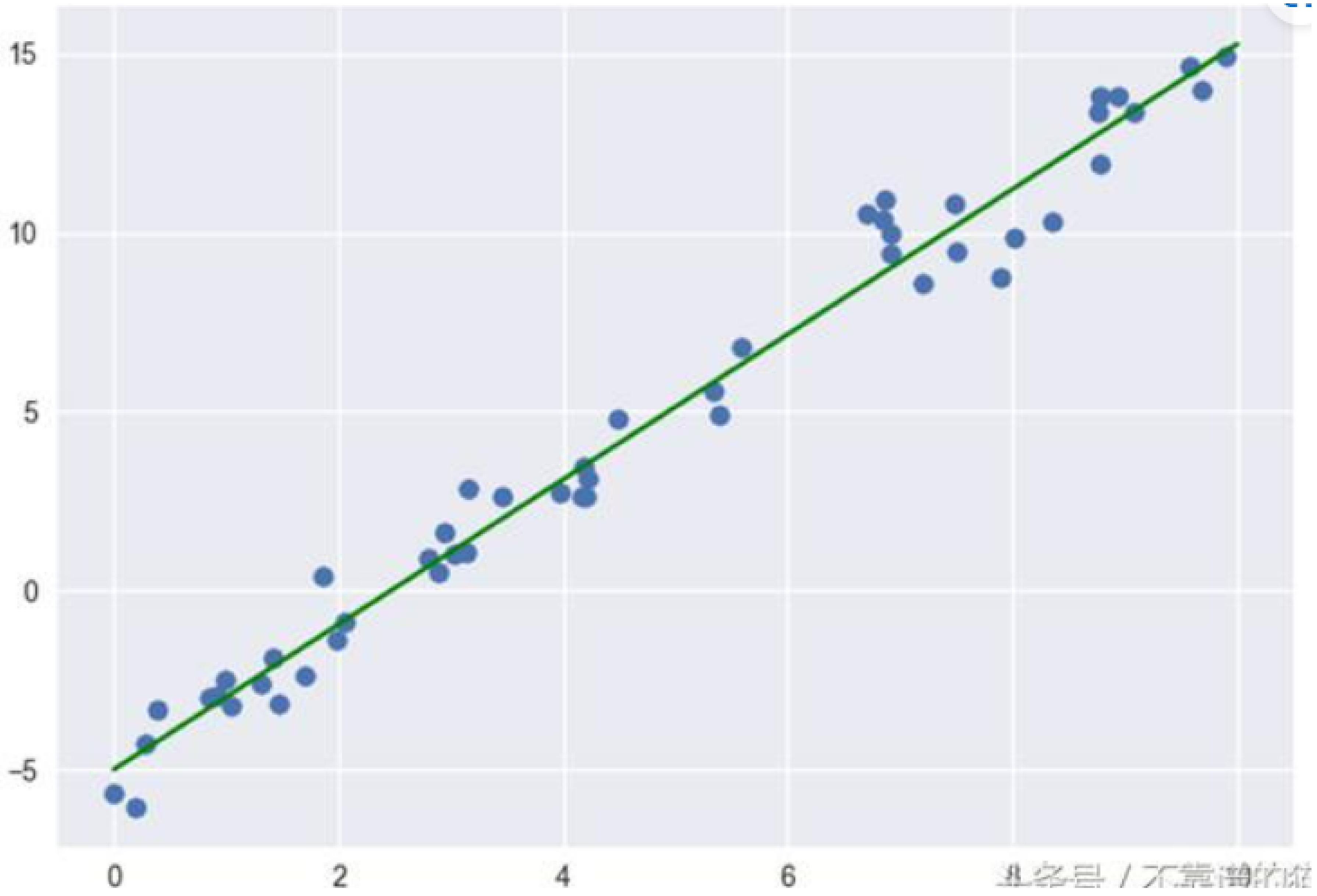
楊豐宇

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

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2 \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

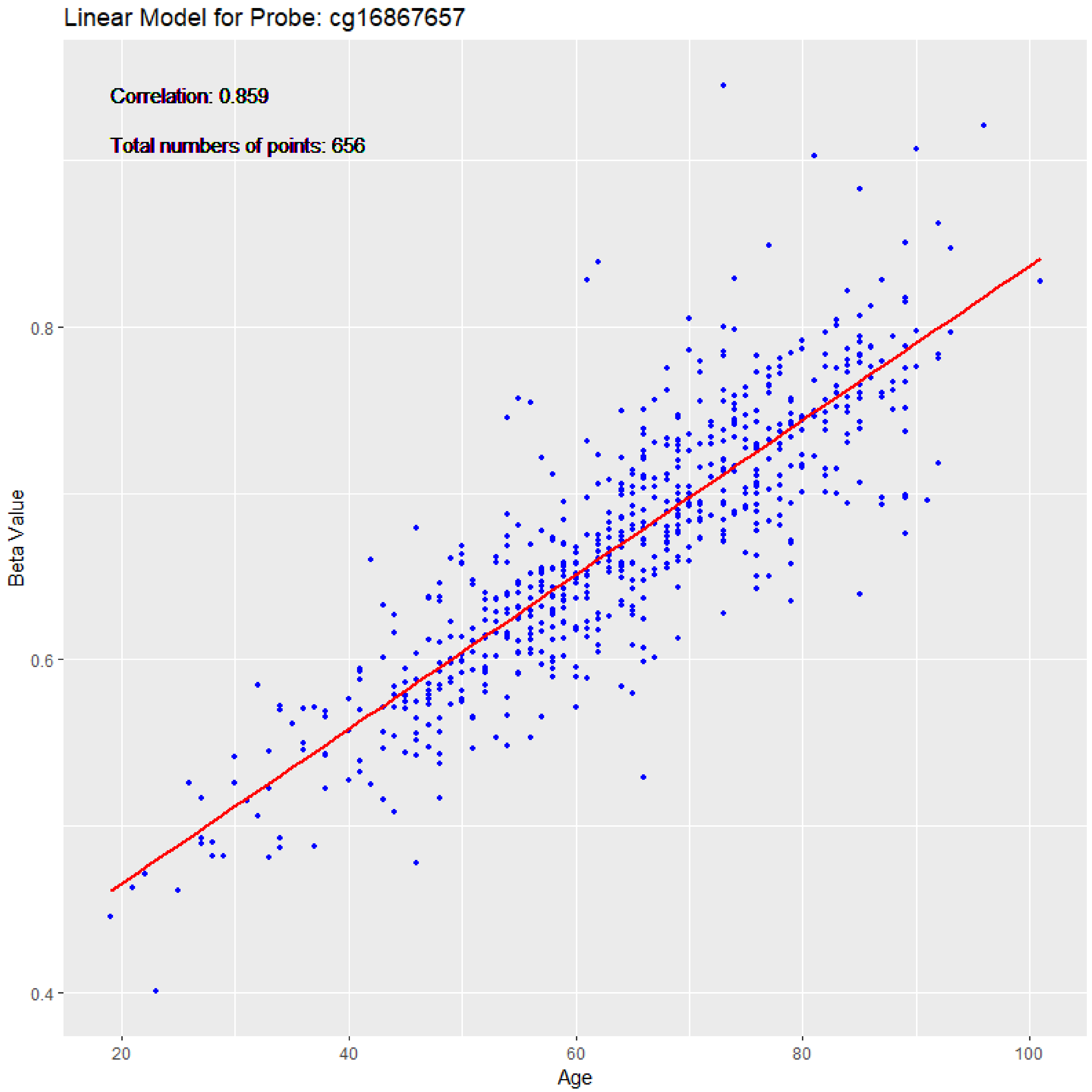
$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$$


15
10
5
0
-5
0
2
4
6
8
10


15
10
5
0
-5
0
2
4
6

Linear Regression in R

Linear Regression in R




```
> summary(model_probe_cg16867657)

Call:
lm(formula = y ~ x)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.150152 -0.022994 -0.002232  0.021933  0.233710 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.3726630  0.0071218   52.33   <2e-16 ***
x           0.0046411  0.0001084   42.82   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04088 on 654 degrees of freedom
Multiple R-squared:  0.7371,	Adjusted R-squared:  0.7367 
F-statistic:  1833 on 1 and 654 DF,  p-value: < 2.2e-16
```

Linear Regression in R

```r
library(limma)
library(MEAL)

# Assuming the probe IDs are listed in the 'probe_id' column of the dataset
probe_ids <- unique(fData(gse40279_matrix)$ID)

# Create an empty list to store the regression results for each probe
probe_regression_results <- list()

for (probe_id in probe_ids) {
  # Select data for the current probe
  probe_data <- subset(gse40279_matrix, fData(gse40279_matrix)$ID == probe_id)

  # Extract the age and beta value
  x <- probe_data$age
  y <- assayData(probe_data)$exprs[1, ]

  # Create and fit the linear regression model
  model <- lm(y ~ x)

  # Store the probe ID and regression model in the results list
  probe_regression_results[[probe_id]] <- model
}
```

Linear Regression in R

Name	Type	Value
probe_regression_results	list [473034]	List of length 473034
cg00000029	list [12] (S3: lm)	List of length 12
cg00000108	list [12] (S3: lm)	List of length 12
cg00000109	list [12] (S3: lm)	List of length 12
cg00000165	list [12] (S3: lm)	List of length 12
cg00000236	list [12] (S3: lm)	List of length 12
cg00000289	list [12] (S3: lm)	List of length 12
cg00000292	list [12] (S3: lm)	List of length 12
cg00000321	list [12] (S3: lm)	List of length 12
cg00000363	list [12] (S3: lm)	List of length 12
cg00000622	list [12] (S3: lm)	List of length 12
cg00000658	list [12] (S3: lm)	List of length 12
cg00000714	list [12] (S3: lm)	List of length 12
cg00000721	list [12] (S3: lm)	List of length 12

probe_regression_results

Linear Regression in R

LR.R | data_of_interest | correlation_df

Filter

	probe_id	correlation_coefficient
301867	cg16867657	0.8585363
125882	cg06639320	0.7471026
422694	cg24724428	0.7445793
386708	cg22454769	0.7439943
412467	cg24079702	0.7074021
143282	cg07553761	0.7000595
374187	cg21572722	0.6871400
128662	cg06784991	0.6734072
94605	cg04875128	0.6650027
266488	cg14692377	0.6541842
390948	cg22736354	0.6452676
143171	cg07547549	0.6312118
154122	cg08160331	0.6287802
52539	cg02650266	0.6284389

Showing 1 to 15 of 473,034 entries, 2 total columns

References:

Montgomery, D. C., Peck, E. A., & Vining, G. G. (1993a). Introduction to linear regression analysis. Journal of the American Statistical Association, 88(421), 383. https://doi.org/10.2307/2290746

O, I., & Guest, P. G. (1961a). Numerical methods of curve fitting. Journal of the American Statistical Association. https://doi.org/10.2307/2282040

Supervised Machine Learning: regression and classification. (n.d.-a). Coursera. https://www.coursera.org/learn/machine-learning?specialization=machine-learning-introduction

What is a complete list of the usual assumptions for linear regression? (n.d.-a). Cross Validated. https://stats.stackexchange.com/questions/16381/what-is-a-complete-list-of-the-usual-assumptions-for-linear-regression

References:

Chwang. (2021a, December 25). Machine Learning — 給自己的機器學習筆記 — Linear Regression — 迴歸模型介紹與公式原理教學. Medium. https://reurl.cc/3xAgYj

iThome. (n.d.-a). [Day 8] 線性迴歸 (Linear Regression) - IT 邦幫忙::一起幫忙解決難題,拯救 IT 人的一天. iT 邦幫忙::一起幫忙解決難題，拯救 IT 人的一天. https://ithelp.ithome.com.tw/articles/10268453

造成上次兩群資料(gse30870&gse40279)所分析出來p-value < 1e-5 的探針們不太一致的可能原因:

1. gse40279 選取的Young-Old 不像gse30870選取的Young-Old那麼極端
2. 不同樣本的差異
3. 取樣本的過程中（保存/提取基因的過程）基因被污染