Regression analysis, a statistical technique for estimating the relationships among variables.

Examples

- 消費～収入
- 總收入～受教育時間
- 地層下陷程度～捷運施工費用
- 經濟指數～保險費用
- 糖尿病患病數～時間趨勢

Another Examples

- 糧食輸出~蝗蟲數目
- 嬰兒出生數~國內生產總值(GDP)
- 生涯收入~基因分數
- 死亡率~平均每天喝牛奶CC數

Variables

Response variables = Y-variables
= output variables
= dependent variables
= outcome variables

Predictor variables = x-variables
= input variables
= independent variables
= explanatory variables
= covariates
= features

Regression model

Response variable = Model function + Random error

e.g. $$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

(A first – order linear regression model)

or $$Y = \beta_0 + \beta_{11} X_1 + \beta_{12} X_1^2 + \varepsilon$$

(A second – order linear regression model)

Poisson regression model, logistic regression model, Cox proportional model,…, etc

Linear regression

$$Y \sim \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \cdots$$

Simple linear regression

$$Y \sim \beta_0 + \beta_1 X$$

R command- lm

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \cdots, n$$

- Condition(A1). $E(\varepsilon_i) = 0, \;\; i = 1, \cdots, n.$
- Condition(A2). $Var(\varepsilon_i) = \sigma^2, \;\; i = 1, \cdots, n.$
- Condition(A3). $Cov(\varepsilon_i, \varepsilon_j) = 0,$ for $i \neq j.$
- Note that: $x_i, \; i = 1, \cdots, n$ not random variables

Random predictor variables

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ i = 1, \cdots, n$$

- Condition(A1'). $E(\varepsilon_i \mid X_i) = 0, \quad i = 1, \cdots, n.$
- Condition(A2'). $Var(\varepsilon_i \mid X_i) = \sigma^2, \quad i = 1, \cdots, n.$
- Condition(A3'). $Cov(\varepsilon_i, \varepsilon_j \mid X_i, X_j) = 0, \text{ for } i \neq j.$

$$E(Y_i \mid X_i = x_i) = \beta_0 + \beta_1 x_i, \quad i = 1, \cdots, n$$

$$= \mu_i$$

$$Var(Y_i \mid X_i = x_i) = \sigma^2.$$

$$Var\Big(E\big(Y \mid X = x\big)\Big) \le Var\big(Y\big) \ ?$$

Estimation of parameters

- Least squares method, LSE
- Maximum likelihood method, MLE

Least Squares Estimation (LSE)

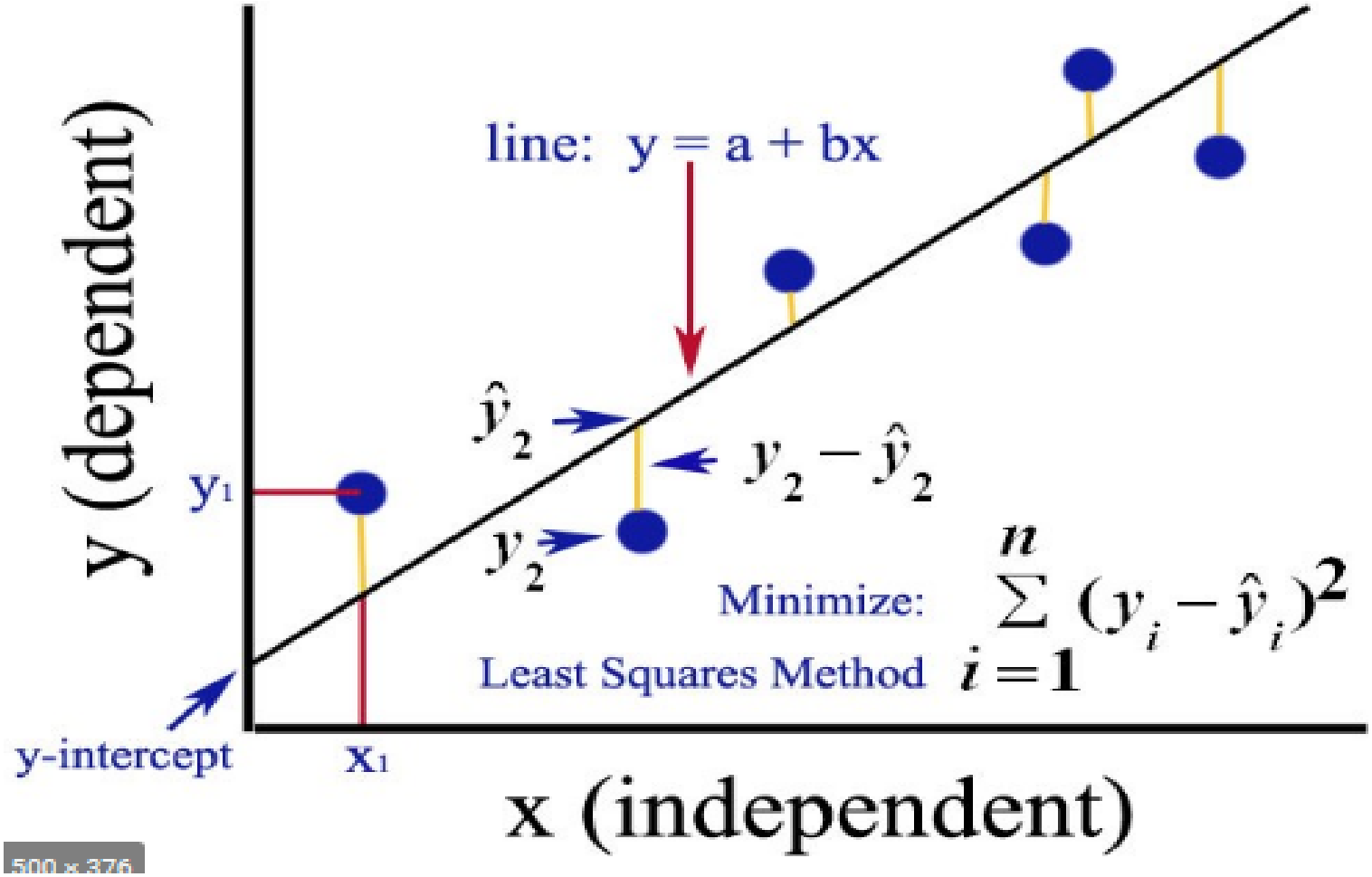
Data : $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$

Model : $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad , i = 1,2,\cdots,n$

LSE of β_0 & β_1 : $\hat{\beta}_0$ & $\hat{\beta}_1$ (Fig. 1.5)*

predicted value of y given x : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

*：Applied Regression Analysis (3rd edition). Norman R. Draper, Harry Smith (1998)



https://www.quora.com/What-is-the-difference-between-linear-regression-and-least-squares

Derivation - LSE

Sum of squares (SS) function :

$$S = S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\text{Find } (\hat{\beta}_0, \hat{\beta}_1) \text{ such that } S(\hat{\beta}_0, \hat{\beta}_1) = \min_{\beta_0, \beta_1} S(\beta_0, \beta_1)$$

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i)$$

Derivation – LSE (cont.)

normal equations（正規方程式，非常態方程式）:

$$\sum_{i=1}^{n}\left(y_i-\widehat{\beta}_0-\widehat{\beta}_1 x_i\right)=0$$

$$\sum_{i=1}^{n} x_i\left(y_i-\widehat{\beta}_0-\widehat{\beta}_1 x_i\right)=0$$

$$\widehat{\beta}_1=\frac{S_{xy}}{S_{xx}}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{\sum_{i=1}^{n}x_iy_i-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)/n}{\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2/n}$$

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

So, $\hat{y}=\widehat{\beta}_0+\widehat{\beta}_1 x=\bar{y}+\widehat{\beta}_1(x-\bar{x})$

$$S^2 = \widehat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2}{n-2} = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2}}{n-2}$$

Then, we have

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = 0, \qquad \sum_{i=1}^{n} x_i \hat{\varepsilon}_i = 0,$$

$$\sum_{i=1}^{n} \widehat{y}_i \hat{\varepsilon}_i = 0.$$

Let $c_i = \dfrac{x_i - \overline{x}}{S_{xx}}$.

Then, we have

$$\widehat{\beta}_1 = \sum_{i=1}^{n} c_i y_i ,$$

$$\widehat{\beta}_0 = \sum_{i=1}^{n} k_i y_i = \sum_{i=1}^{n} \left(\frac{1}{n} - \overline{x} c_i \right) y_i .$$

$$\sum_{i=1}^{n} c_i = 0,$$

$$\sum_{i=1}^{n} c_i x_i = 1,$$

$$\sum_{i=1}^{n} c_i^2 = \frac{1}{S_{xx}}.$$

Random predictor variables

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Then, we have $Cov(X, Y) = \beta_1 Var(X)$,

$$\beta_1 = \frac{Cov(X, Y)}{Var(X)}.$$

Naturally, we also estimate the slope

$$\widehat{\beta}_1 = \frac{\widehat{Cov}(X, Y)}{\widehat{Var}(X)}$$

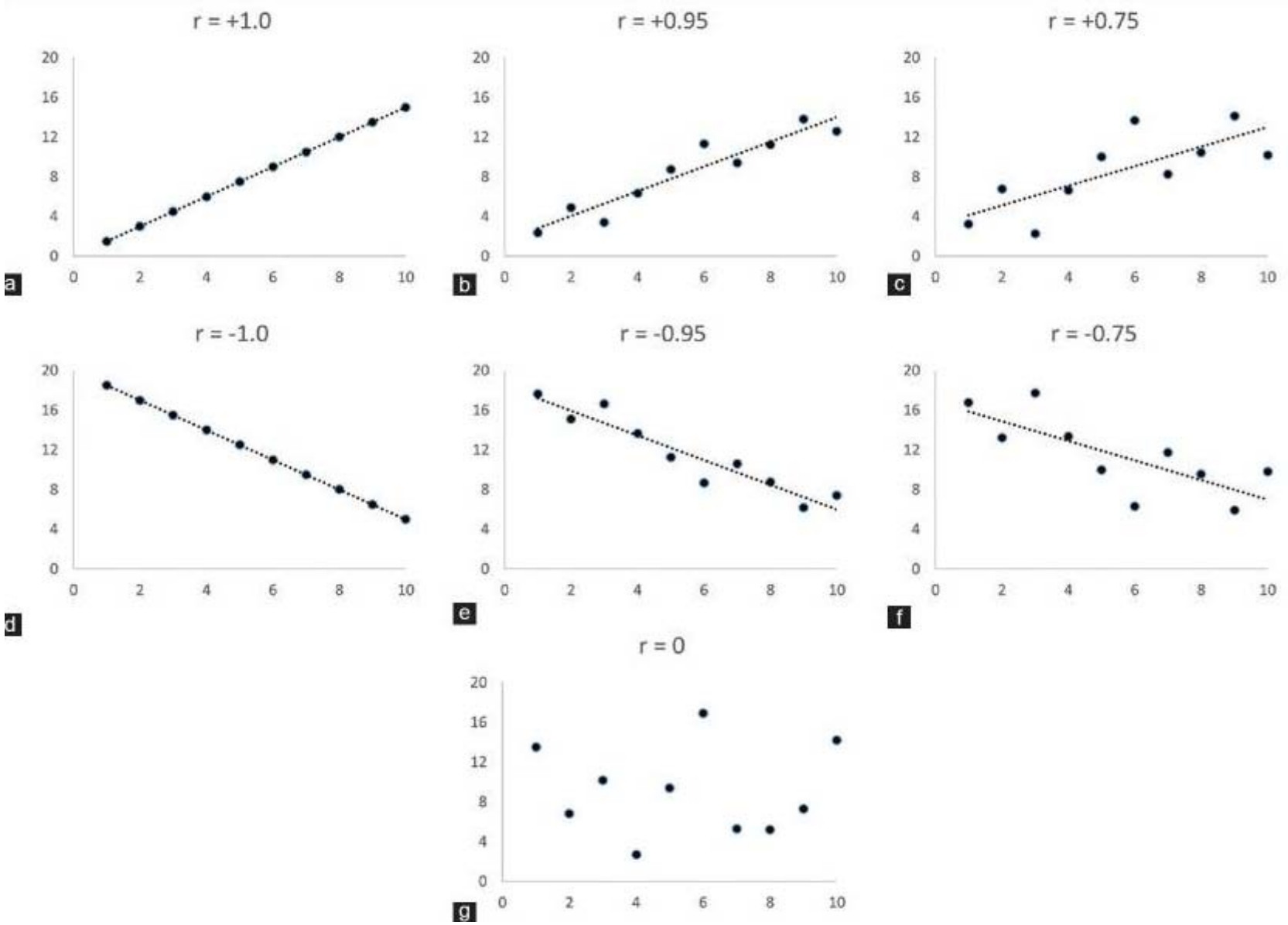
$$= \frac{\widehat{sd}(Y)}{\widehat{sd}(X)} \cdot \widehat{corr}(X, Y).$$

(Pearson) Correlation coefficient

$$corr(X,Y) = \rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

$$-1 \leq \rho_{X,Y} \leq 1$$

A correlation of 0 shows no linear relationship between the movement of two variables.



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5079093/

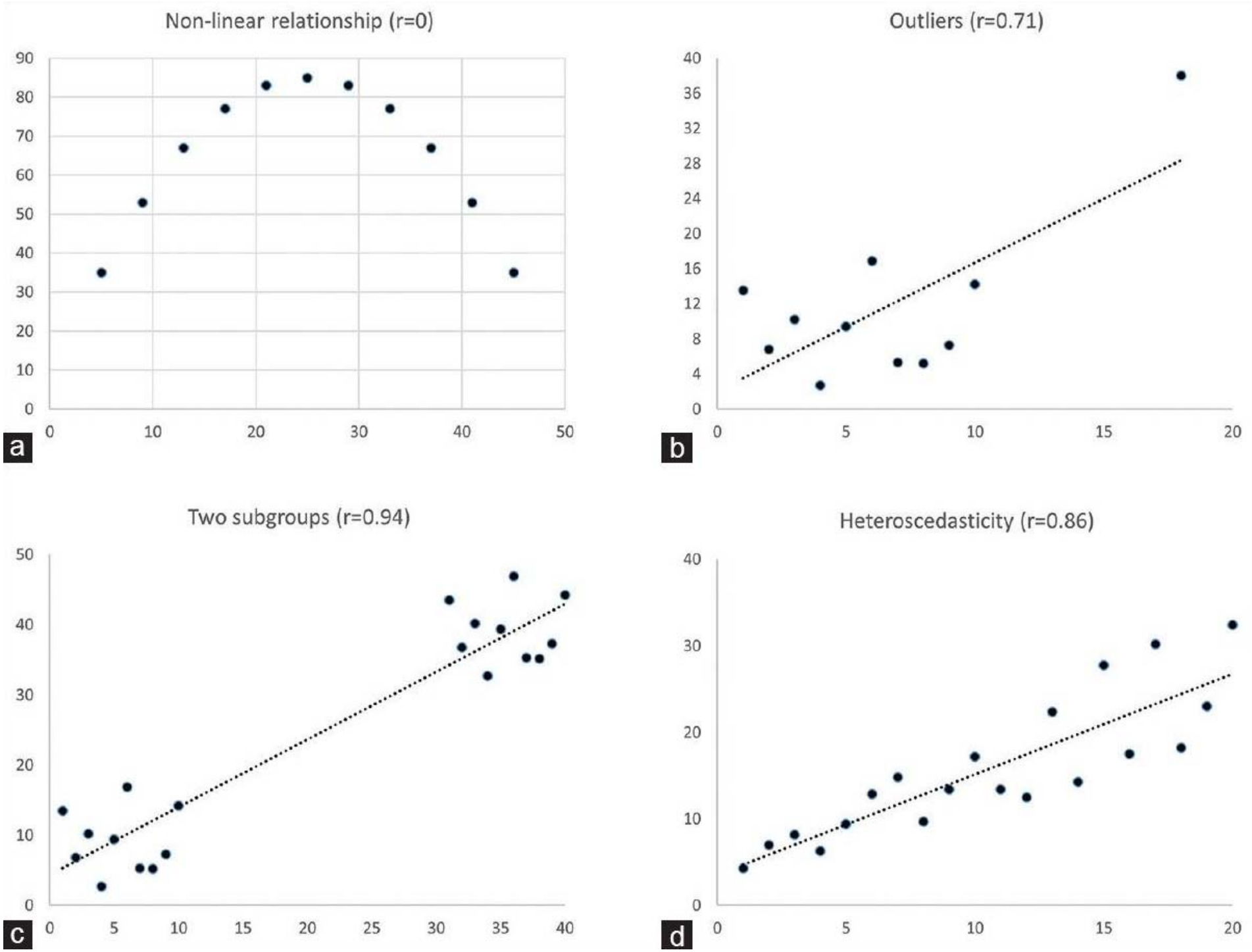
$$\widehat{corr}(X, Y) = r_{X,Y} = \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}}$$

under $H_0 : \rho = 0$ and the normal assumption on errors,

$$\frac{r}{\sqrt{1 - r^2} / \sqrt{n - 2}} \sim t_{n-2}$$

- With medium- to large-sized samples, these methods show even small correlation coefficients to be highly significant and hence their use is generally eschewed.

- With very small sample size (say 3–6 observations), a relationship may appear to be present even though none exists.


Non-linear relationship (r=0)
a
Outliers (r=0.71)
b
Two subgroups (r=0.94)
c
Heteroscedasticity (r=0.86)
d

WHEN SHOULD CORRELATION NOT BE USED?

- Nonlinear relationship
- Outliers
- Two subgroups
- Heteroscedasticity

Correlation does not imply causation.

- 各國每人平均巧克力消費量~各國每千萬人口諾貝爾得獎者人數
- 各國每週食用冰淇淋的數量~各國游泳池中溺水事件的件數

https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation

Expected values of LSEs

$$E\left(\widehat{\beta}_1\right) = \beta_1$$

$$E\left(\widehat{\beta}_0\right) = \beta_0$$

Variances and covariance of LSEs

$$Var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{S_{xx}}$$

$$Var\left(\hat{\beta}_0\right) = \sum_{i=1}^{n} k_i^2 \sigma^2, k_i = \frac{1}{n} - \bar{x} c_i$$

$$= \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \sigma^2$$

$$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\frac{\bar{x}\sigma^2}{S_{xx}}$$

More on LSE

Note : 1. The LS line goes through $(\bar{X}, \bar{Y})$

2. Remove β_0 from the model by centering the data

$$Y - \bar{Y} = \beta_0 + \beta_1 X + \varepsilon - \bar{Y}$$

$$= (\beta_0 + \beta_1 \bar{X} - \bar{Y}) + \beta_1 (X - \bar{X}) + \varepsilon$$

$$\Rightarrow \quad y = \beta_0' + \beta_1 x + \varepsilon ,$$

where $y = Y - \bar{Y}$, $\beta_0' = \beta_0 + \beta_1 \bar{X} - \bar{Y}$ and $x' = X - \bar{X}$

LS estimates of β_0' is $b_0' = \bar{y} - b_1 \bar{x} = 0$, since $\bar{x} = \bar{y} = 0$

So, the centered model is

$$Y - \bar{Y} = \beta_1 (X - \bar{X}) + \varepsilon$$

BLUE and Gauss-Markov Theorem

- LSEs are the best linear unbiased estimator (BLUE). "best": the lowest variance of the estimate, as compared to other unbiased, linear estimators (Gauss-Markov Theorem). i.e.: for any linear unbiased estimators $\tilde{\beta}_1$,

$$Var\left(\widehat{\beta}_1\right) \leq Var\left(\tilde{\beta}_1\right).$$

More on estimates of residuals

Data : $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$

Model : $y_i = \beta_1 x_i + \varepsilon_i \quad , i = 1, 2, \cdots, n$

$$\sum_{i=1}^{n} x_i \hat{\varepsilon}_i = \sum_{i=1}^{n} x_i \left(y_i - \hat{\beta}_1 x_i \right) = 0,$$

$$\sum_{i=1}^{n} \hat{\varepsilon}_i \neq 0, \qquad \sum_{i=1}^{n} \hat{y}_i \hat{\varepsilon}_i \neq 0.$$

Proportions

$$E\left(\hat{Y}_i\right) = \beta_0 + \beta_1 x_i, \qquad Var\left(\hat{Y}_i\right) = h_{ii}\sigma^2,$$

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$E\left(\hat{\varepsilon}_i\right) = 0, \qquad Var\left(\hat{\varepsilon}_i\right) = \left(1 - h_{ii}\right)\sigma^2,$$

$$Cov\left(Y_i, \hat{Y}_j\right) = h_{ij}\sigma^2 = Cov\left(\hat{Y}_i, \hat{Y}_j\right),$$

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

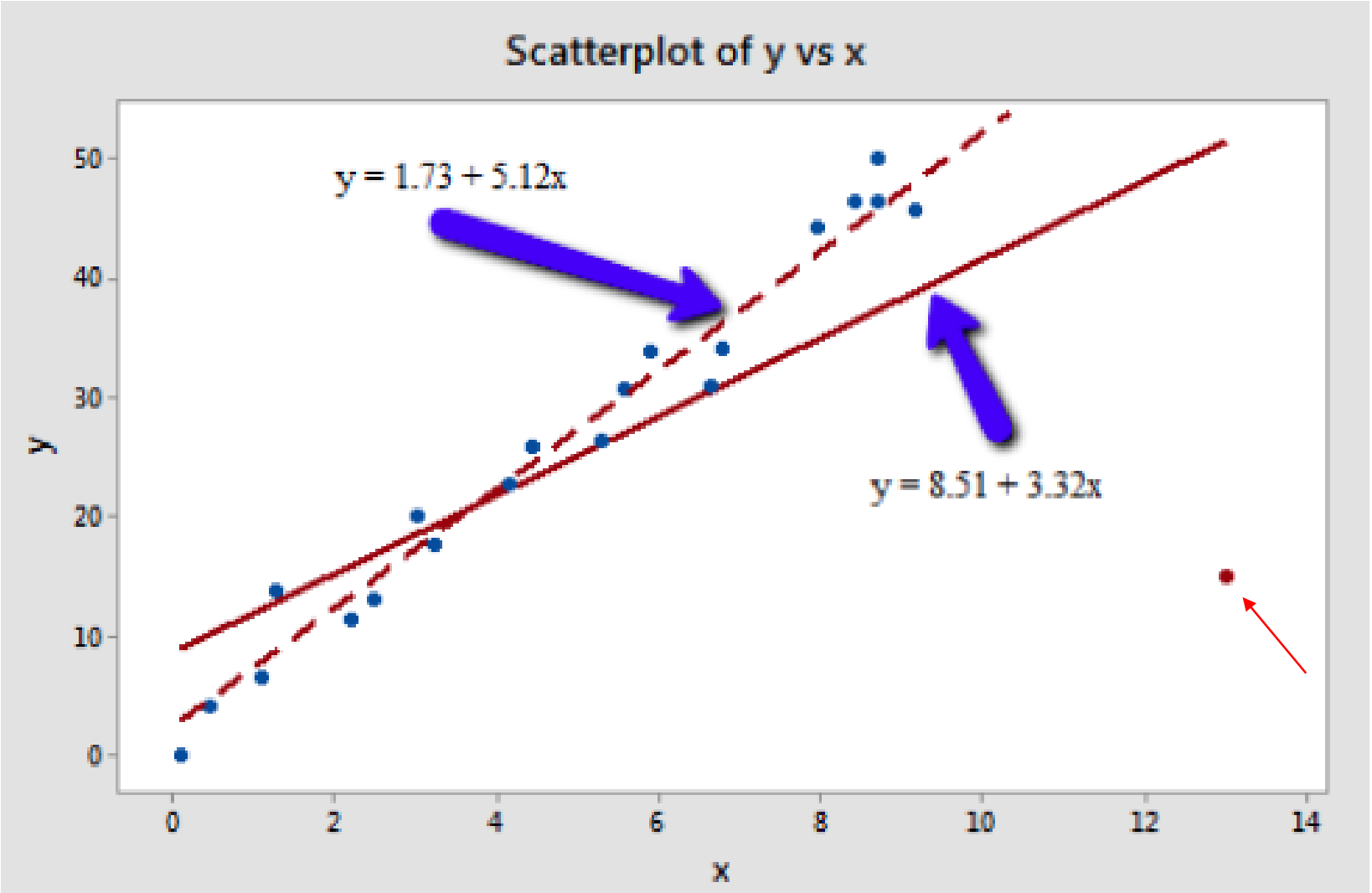
$$Cov\left(\hat{\varepsilon}_i, \hat{\varepsilon}_j\right) = -h_{ij}\sigma^2, i \neq j,$$

$$Cov\left(Y_i, \hat{\varepsilon}_i\right) = \left(1 - h_{ii}\right)\sigma^2,$$

$$Cov\left(Y_i, \hat{\varepsilon}_j\right) = -h_{ij}\sigma^2, i \neq j,$$

$$Cov\left(\hat{Y}_i, \hat{\varepsilon}_j\right) = 0, i \neq j.$$

h_{ii} is called the leverage of i-th observations.



https://online.stat.psu.edu/stat462/node/170/

Leverage

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \in \left[\frac{1}{n}, 1\right],$$

- 越接近變數平均越接近0；反之，越接近1
- 只和變數有關，無關反應變數
- 較大值容易是影響點 (influential point)，但不必然
- *R* command- hatvalues, rstudent, boxplot, diffits, cooks.distance (如何判斷離群值)

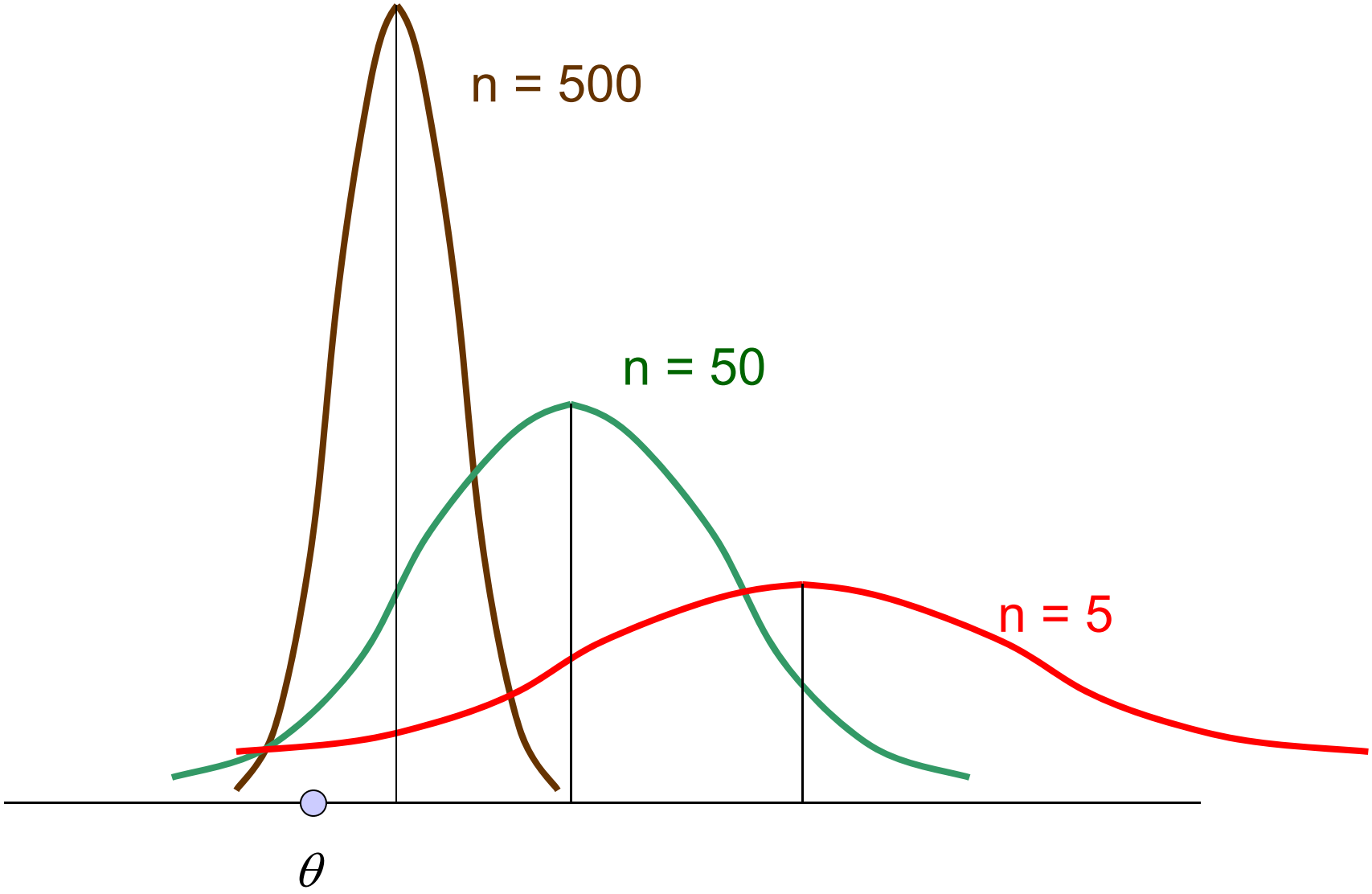
Properties of a good estimator

- Unbiasedness $E(\hat{\theta}) = \theta$
- Efficiency $\hat{\theta} = \arg_\theta MSE\left(\hat{\theta}\right) = \arg_\theta E[(\hat{\theta} - \theta)^2]$

$= \arg_\theta \left(Var\left(\hat{\theta}\right) + Bias^2\left(\hat{\theta}\right)\right)$

- Sufficiency $\hat{\theta}$ uses all the information about the population parameter that the sample can provide; Neymann-Fisher factorization theorem
- Consistency $\hat{\theta} \xrightarrow{p} \theta$
- Completeness…

Example of a consistent estimator



Properties of MLE

- Consistency
- Asymptotic normality

$$\sqrt{n}\left(\hat{\theta}-\theta\right)\xrightarrow{d} N\left(0,\frac{1}{I\left(\theta\right)}\right),$$

$$I\left(\theta\right)=-E\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta\left(Y\right)\right]$$

- Invariance principle

Let $\tau = g(\theta)$ be a function of θ. Let $\hat{\theta}_n$ be an MLE of θ. Then $\hat{\tau}_n = g(\hat{\theta}_n)$ is an MLE of τ.

- Cramer-Rao Lower Bound (CRLB)

CRLB describes a lower bound on the variance of estimators of the deterministic parameter θ . For unbiased estimator $\hat{\theta}(Y)$, it can be as

$$\mathrm{Var}\left(\widehat{\theta}(Y)\right) \geqslant \frac{1}{I(\theta)} .$$

Linear combination of independent normal random variables

If $X_1, X_2, \ldots, X_n$ are mutually independent normal variables with means $\mu_1, \mu_2, \ldots, \mu_n$ and variables $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, then

$$Y = \sum_{i=1}^{n} c_i X_i \sim N\left(\sum_{i=1}^{n} c_i \mu_i, \sum_{i=1}^{n} c_i^2 \sigma_i^2 \right).$$

Normal assumptions

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ i = 1, \cdots, n$$

- Condition(B1). $\varepsilon_i \sim_{iid} N\left(0, \sigma^2\right).$
- Condition(B1'). $\varepsilon_i \mid X_i = x_i \sim_{iid} N\left(0, \sigma^2\right).$

Maximum likelihood method (MLE)

Data $: (x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$

Model $: y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad , i = 1,2,\cdots,n$

Under Condition (B1), the MLEs of

$$\left(\beta_0, \beta_1, \sigma^2\right) = \left(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2\right).$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right),$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right),$$

$$E\left(\hat{\sigma}^2\right) = \sigma^2, \qquad \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2),$$

$$\hat{\beta}_0 \amalg \hat{\sigma}^2, \qquad \hat{\beta}_1 \amalg \hat{\sigma}^2.$$ Cochran's Theorem

Confidence interval and prediction

$$\frac{\hat{\beta}_1 - \beta_1}{S\sqrt{\frac{1}{S_{xx}}}} \sim t(n-2),$$

$$\frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}} \sim t(n-2),$$

$$100(1-\alpha)\%\ \text{C.I. for } \beta_1 = \left(\hat{\beta}_1 \mp \frac{s}{\sqrt{S_{xx}}} t_{\alpha/2,(n-2)} \right),$$

$$100(1-\alpha)\%\ \text{C.I. for } \beta_0 = \left(\hat{\beta}_0 \mp s\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}} t_{\alpha/2,(n-2)} \right),$$

$$100(1-\alpha)\%\ \text{C.I. for } \sigma^2 =$$

$$\left(\frac{(n-2)s^2}{\chi^2_{\alpha/2,(n-2)}}, \frac{(n-2)s^2}{\chi^2_{1-\alpha/2,(n-2)}} \right),$$

$$\frac{\hat{Y}^{*} - E\left(Y \mid X = x^{*}\right)}{S\sqrt{\frac{1}{n} + \frac{\left(x^{*} - \bar{x}\right)^{2}}{S_{xx}}}} \sim t\left(n-2\right)$$

$$\hat{Y}^{*} = \hat{E}\left(Y \mid X = x^{*}\right) = \hat{\beta}_{0} + \hat{\beta}_{1}x^{*}$$

$$E\left(Y \mid X = x^{*}\right) = \beta_{0} + \beta_{1}x^{*}$$

$$100(1-\alpha)\%\ \text{C.I. for } E(Y \mid X = x^*) =$$

$$\left(\hat{y}^* \mp s\sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}} t_{\alpha/2,(n-2)} \right),$$

New observations (x_{n+1}, Y_{n+1}) and prediction interval

$$E\left(\hat{Y}_{n+1} - Y_{n+1}\right) = 0,$$

$$Var\left(\hat{Y}_{n+1} - Y_{n+1}\right) = \sigma^2\left[1 + \frac{1}{n} + \frac{\left(x_{n+1} - \bar{x}\right)^2}{S_{xx}}\right],$$

$$\hat{Y}_{n+1} - Y_{n+1} \sim N\left(0, \left[1 + \frac{1}{n} + \frac{\left(x_{n+1} - \bar{x}\right)^2}{S_{xx}}\right]\sigma^2\right),$$

$$\frac{\hat{Y}_{n+1} - Y_{n+1}}{S\sqrt{1 + \frac{1}{n} + \frac{\left(x_{n+1} - \bar{x}\right)^2}{S_{xx}}}} \sim t\left(n-2\right),$$

$100\left(1-\alpha\right)\%$ predicted interval for $Y_{n+1} =$

$$\left(\hat{y}_{n+1} \mp s\sqrt{1 + \frac{1}{n} + \frac{\left(x_{n+1} - \bar{x}\right)^2}{S_{xx}}} t_{\alpha/2,(n-2)} \right).$$

Analysis of variance (ANOVA) for linear regression model

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

$$SST = SSR + SSE$$

- SST: total sum of squares, total sample variability, total

- SSR: regression sum of squares, variability explained by the model, (between, treatment, group, factor,) SSreg

$$SSR = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{S_{xy}^2}{S_{xx}}$$

- SSE: residual sum of squares, unexplained variability, (within,) error, SSres, RSS

$$SSE = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 - \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

Thus,

$$\frac{SST}{\sigma^2} = \frac{SSR}{\sigma^2} + \frac{SSE}{\sigma^2}$$

- To test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ under H_0

$$\frac{SST}{\sigma^2} \sim \chi^2_{\text{n-1}}, \frac{SSR}{\sigma^2} \sim \chi^2_1, \frac{SSE}{\sigma^2} \sim \chi^2_{\text{n-2}}$$

$$\text{F} = \frac{SSR}{SSE/(n-2)} \sim F_{1,n-2}$$

ANOVA table

Source of variation (Source)	Sum of squares (SS)	Degrees of freedom (df)	Mean square (MS)	F value (F)	P-value
Regression	SSR	1	$MSR = SSR / 1$	$F = MSR / MSE$	$P\left(F_{1,n-2} > F\right)$
Error	SSE	n-2	$MSE = SSE / \left(n-2\right)$		
Total	SST	n-1			

Coefficient of determination, R^2

$$R^2 = \frac{SSR}{SST} = \frac{\text{Variability explained by the model}}{\text{Total sample variability}}$$

$$= 1 - \frac{SSE}{SST}$$

Understanding

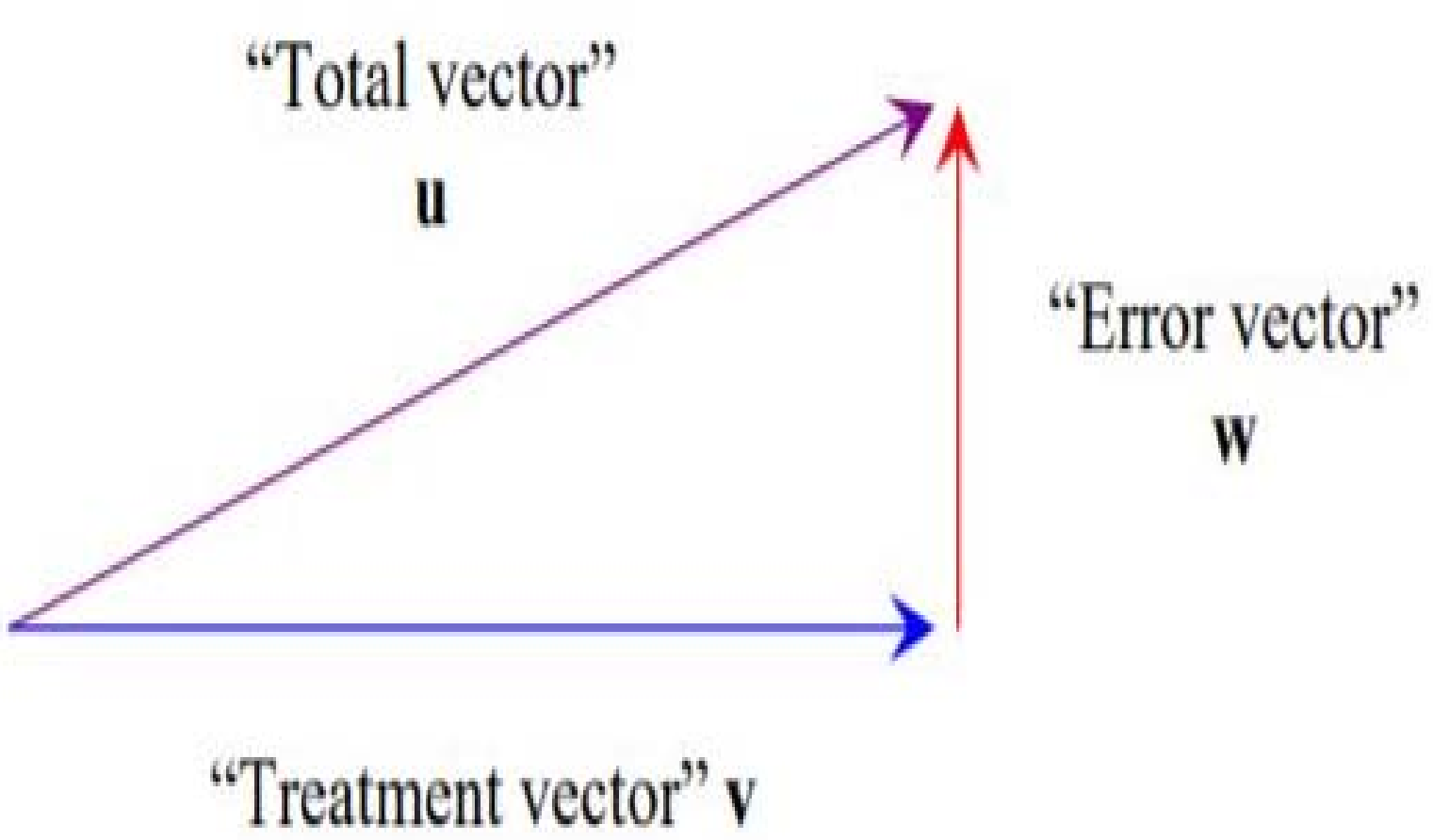
- explain how much variability, as a value between 0 and 1
- the closer that value is to 1: all the data points that are scattered around the graph
- even the value is large, there may be no statistical significance of the explanatory variables in a model (ex. S&P500 ~ cheese (0.75) + sheep(0.99))

http://greenhornfinancefootnote.blogspot.com/2010/08/stupid-data-miner-tricks.html

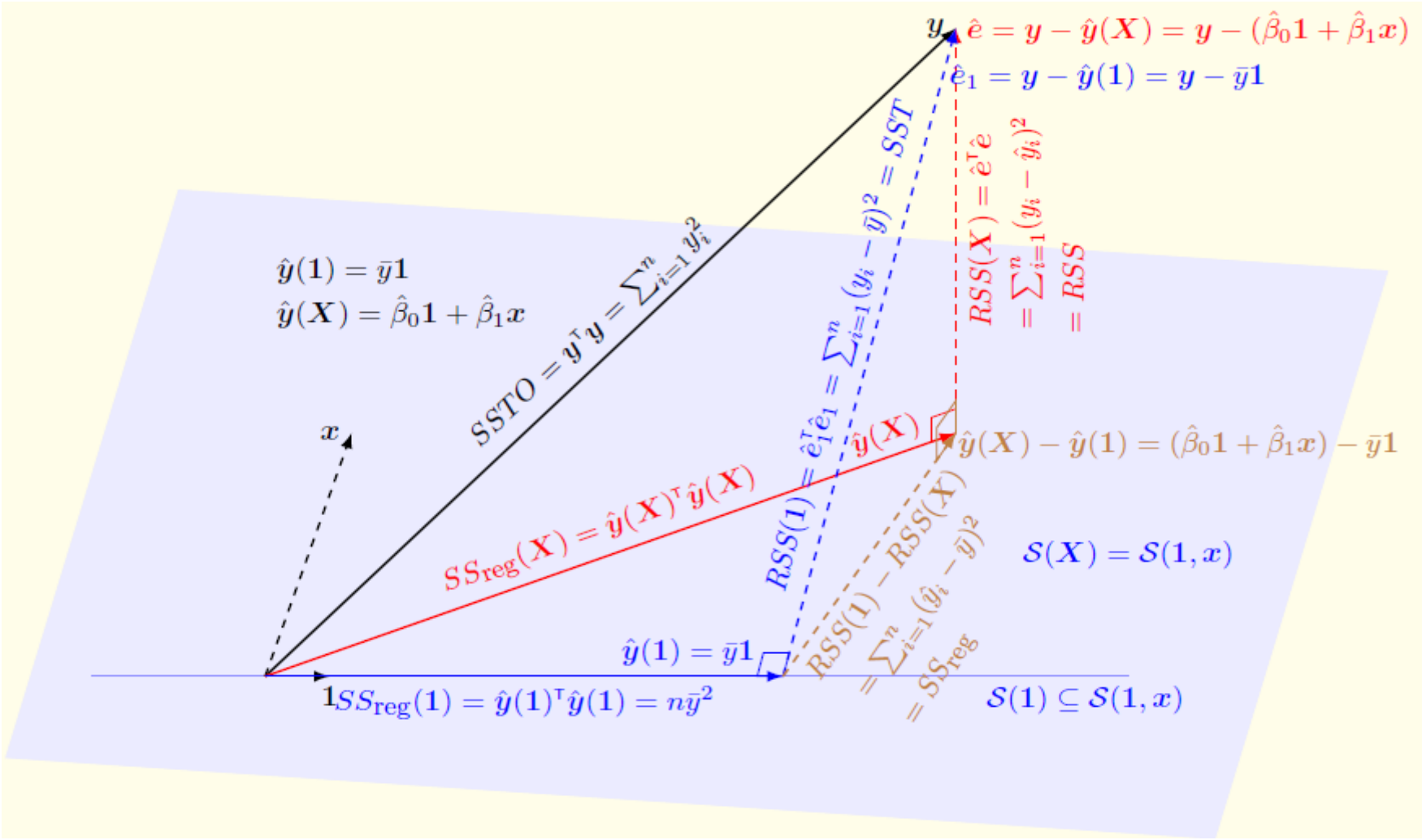
- $R^2 = r_{x,y}^2$ and $\dfrac{r}{\sqrt{1-r^2}\,/\,\sqrt{n-2}} = \dfrac{\hat{\beta}_1}{S\sqrt{\dfrac{1}{S_{xx}}}} \sim t_{n-2}$
- Adjusted coefficient of determination, adj- R^2

$$\text{adj-}R^2 = 1 - \frac{SSE/(n-1-p)}{SST/(n-1)}$$

- adj- R^2 = R^2 on simple regression model



https://stats.stackexchange.com/questions/450160/geometric-interpretation-of-the-difference-between-the-means-anova



https://www.feam.scu.edu.tw/faculty_personal/2