上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

SJTU School of Biomedical Engineering

# 生物医学统计概论(BI148-2)

## Fundamentals of Biomedical Statistics

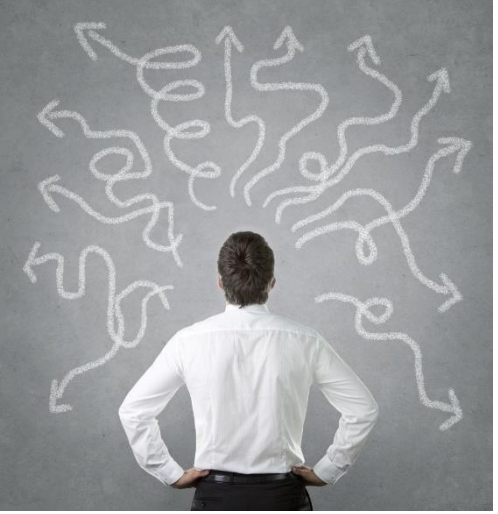
授课：林关宁

2022 春季

# 课程内容安排

| 上课日期 | 章节 | 教学内容 | 教学要点 | 作业 | 随堂测 | 学时 |
|---|---|---|---|---|---|---|
| 2.16 | 1 | 数据可视化，描述性统计 | 1. 课程介绍 & 数据类型 | 作业1 (8%) | 测试1 (8%) | 2 |
| 2.23 | | | 2. 描述性统计Descriptive Statistics & 数据常用可视化 | | | 2 |
| 3.2 | | | 3. 大数定理 & 中心极限定理 | | | 2 |
| 3.9 | | | 4. 常用概率分布 | | | 2 |
| 3.16 | 2 | 推断性统计，均值差异检验 | 5. 统计推断基础-1：置信区间 Confidence Interval * | 作业2 (10%) | 测试2 (10%) | 2 |
| 3.23 | | | 6. 统计推断基础-2：假设检验，I及II类错误，统计量，p-值 | | | 2 |
| 3.30 | | | 7. 数值数据的均值比较-1： 单样本及双样本t-检验，效应量，功效 | | | 2 |
| 4.6 | | | 8. 数值数据的均值比较-2：One-Way ANOVA，正态性检验 | | | 2 |
| 4.13 | | | 9. 数值数据的均值比较-3：Two-Way ANOVA | | | 2 |
| 4.20 | 3 | 比例差异检验 | 10. 样本和置信区间预估 * | 作业3 (6%) | 测试3 (6%) | 2 |
| 4.27 | | | 11. 类别数据的比例比较-1：联立表的卡方检验 | | | 2 |
| 5.11 | | | 13. 类别数据的比例比较-2： 联立表的RR，OR | | | 2 |
| 5.18 | 4 | 协方差，相关分析，回归分析 | 14. 相关分析（Pearson r, Spearman rho, Kendal's tau) * | 作业4 (6%) | 测试4 (6%) | 2 |
| 5.25 | | | 15. 简单回归分析 | | | 2 |
| 6.1 | | | 16. 多元回归Multiple Regression | | | 2 |
| | 5 | Course Summary | 17. 课程总结 * | | | 2 |
| | | | Total | 30% | 30% | 32 |

*** 随堂测试**



# 课程知识点导图

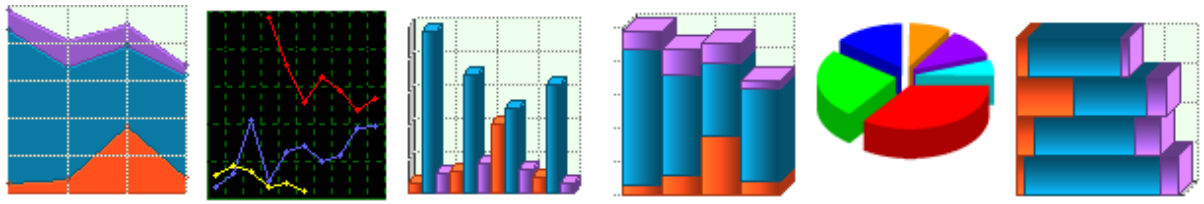
DATA

Data collection
- Epidemiological studies
  - Bias, confounding
- Sample size calculations
- Randomized trials

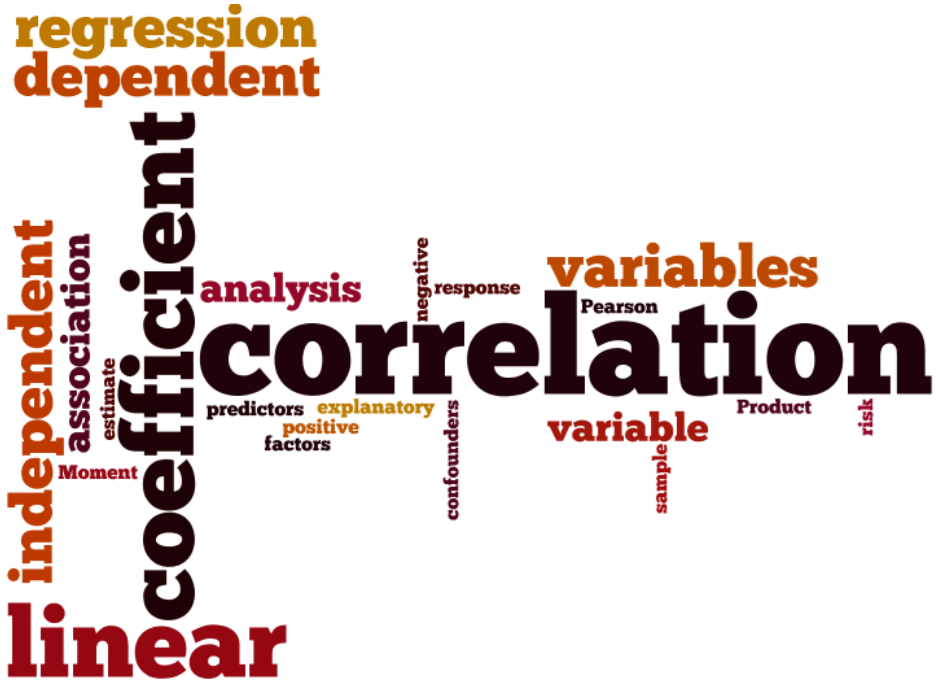
Description
- Tables
- Diagrams

Inference
- Estimation
  - Conf. int.
- Hyp. testing
  - *P*-values
- Models
  - Probability theory, distributions
  - Checking assumptions
- Numerical data
  - 1 group
    - Paired *t*-test
    - Wcx signed ranks test
    - Single-sample *t*-test
  - 2 related groups
  - 2 unrelated groups
    - Two-sample *t*-test
    - Mann-Whitney
  - 3+ groups
    - ANOVA
    - Kruskal-Wallis
  - *x* and *y*
    - Correlation
      - Pearson
      - Spearman
    - Linear regression
      - Estimation
      - Prediction
      - Simple (one *x*)
      - Multiple (many *x*)
- Survival data
  - Kaplan–Meier
  - Log-rank test
  - Cox model
- Logistic regression (*y* dichotomous)
- Categorical data
  - Unrelated
    - $\chi^2$ test
    - Fisher's exact test
  - Related
    - McNemar's test
  - ROC curve
  - Diagnostic tests

单元1
单元2,3
单元2
单元1
单元4
单元3



# Unit 4: Covariance, Correlation & Regression




# 回顾：离散趋势

- 均值（Mean）

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- 方差（Variance）

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

- 标准差 (SD, STD)

$$S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}}$$

- 协方差（covariance）

$$Cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)] = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

# 协方差计算

约翰是个投资者。他的投资组合主要跟踪 S&P 500 标准普尔500指数的表现，约翰想增加ABC公司的股票。在将股票添加到他的投资组合中之前，他想评估股票和标准普尔500指数之间的方向关系

1.获取数据。
首先，约翰获得了ABC公司股票和标准普尔500指数的数据。获得的价格汇总如下表所示

| | S&P 500 | ABC Corp. |
|---|---|---|
| 2013 | 1,692 | 68 |
| 2014 | 1,978 | 102 |
| 2015 | 1,884 | 110 |
| 2016 | 2,151 | 112 |
| 2017 | 2,519 | 154 |

2.计算每项资产的平均（平均）价格

$$\text{Mean (S\&P 500)} = \frac{1{,}692 + 1{,}978 + 1{,}884 + 2{,}151 + 2{,}519}{5} = 2{,}044.80$$

$$\text{Mean (ABC Corp.)} = \frac{68 + 102 + 110 + 112 + 154}{5} = 109.20$$

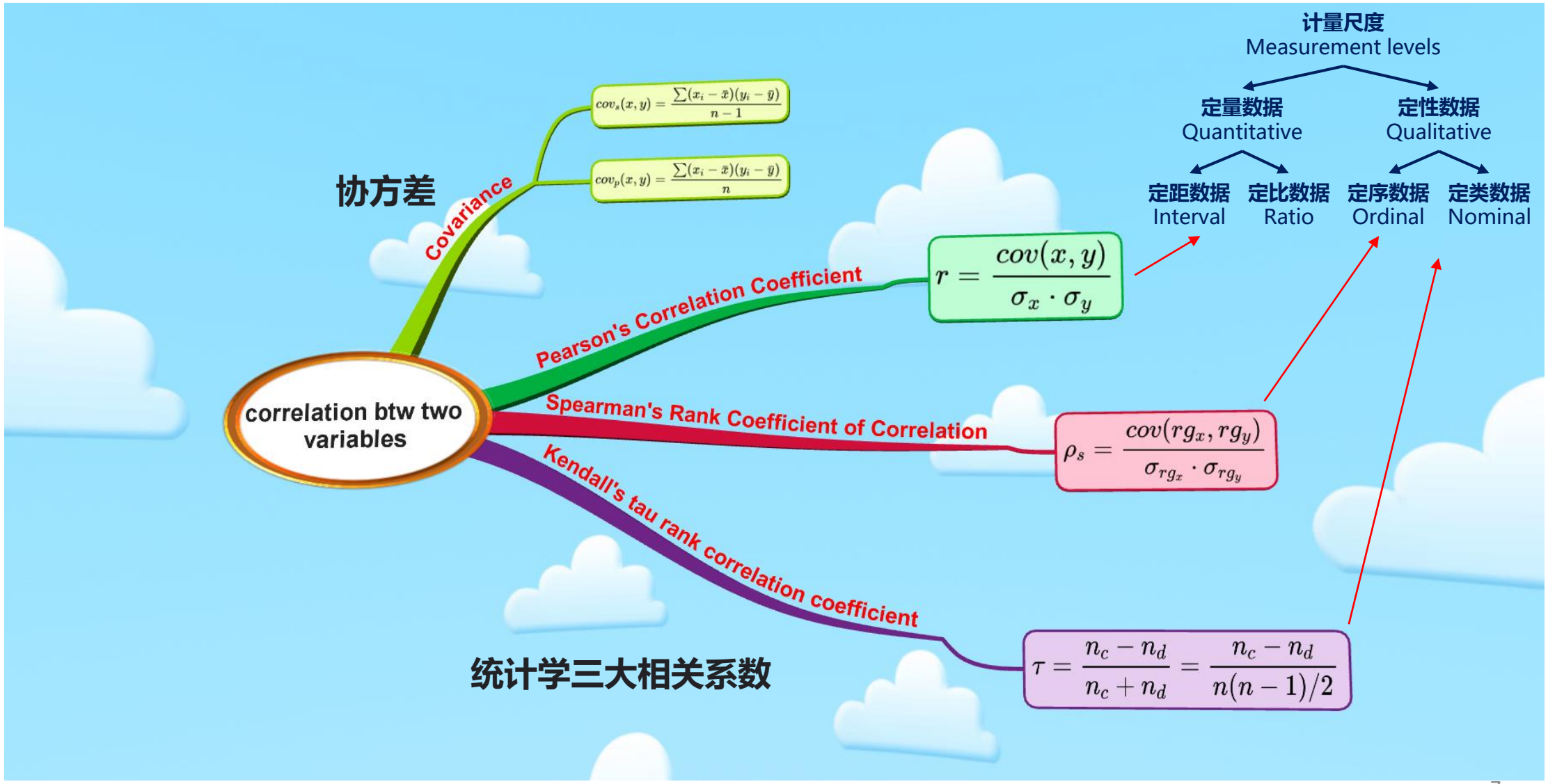
3.对于每种证券，找出每种价值和平均价格之间的差异。

| | | | Step 3 | | Step 4 |
|---|---|---|---|---|---|
| | S&P 500 | ABC Corp. | a | b | a x b |
| 2013 | 1,692 | 68 | -352.80 | -41.20 | 14,535.36 |
| 2014 | 1,978 | 102 | -66.80 | -7.20 | 480.96 |
| 2015 | 1,884 | 110 | -160.80 | 0.80 | -128.64 |
| 2016 | 2,151 | 112 | 106.20 | 2.80 | 297.36 |
| 2017 | 2,519 | 154 | 474.20 | 44.80 | 21,244.16 |
| **Mean** | **2,044.80** | **109.20** | **Sum** | | **36,429.20** |

4.将上一步得到的结果相乘。
5.使用步骤4中计算的数字，找到协方差

$$\text{Cov(S\&P 500, ABC Corp.)} = \frac{36{,}429.20}{5 - 1} = 9{,}107.30$$

# Describing the correlation between two variables

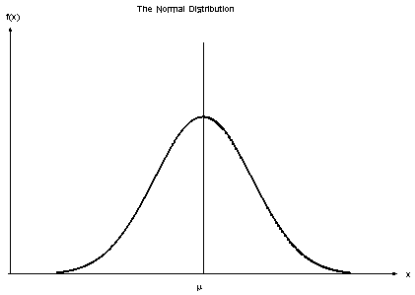


# 三大相关系数之- 皮尔逊相关性系数（Pearson’s correlation）

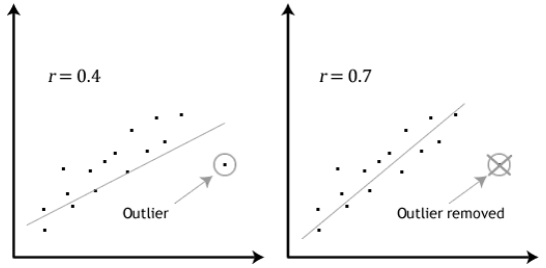
皮尔逊相关也称为积差相关（或积矩相关）是英国统计学家皮尔逊于20世纪提出的一种计算直线相关的方法

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

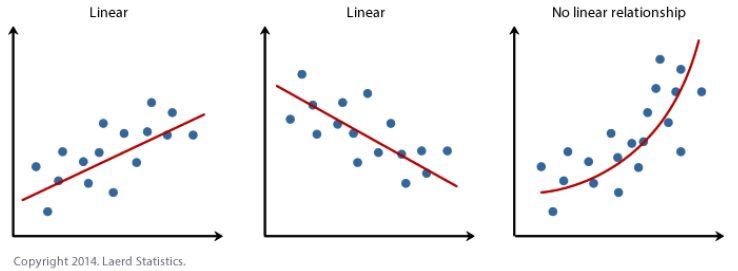
（1）



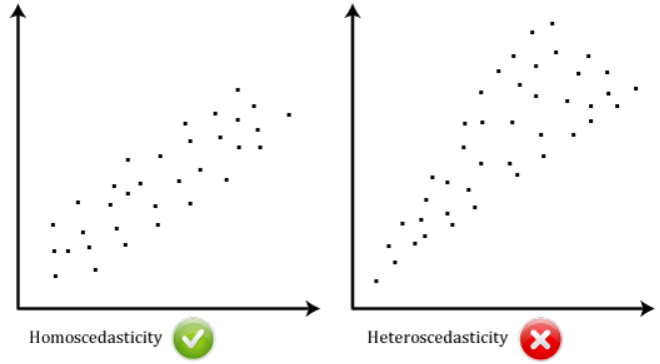

（2）




（4）




（6）




## 协方差与相关性

- 协方差和相关性都主要评估变量之间的关系。与它们之间的关系最接近的类比是方差和标准差之间的关系
- 协方差测量两个随机变量与其期望值之间的总变化。使用协方差，我们**只能测量关系的方向**（变量是否倾向于串联移动或显示反向关系）。然而，**它并没有表明关系的强度，也没有表明变量之间的依赖性**
- 另一方面，相关性**衡量变量之间关系的强度**。相关性是协方差的标度度量。它是无量纲的。换句话说，相关系数始终是一个纯值，不以任何单位测量（no measurement unit)

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

# 三大相关系数之- 皮尔逊相关性系数（Pearson’s correlation）

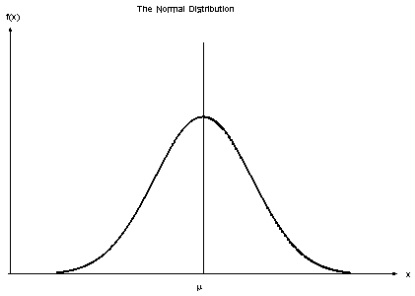
皮尔逊相关也称为积差相关（或积矩相关）是英国统计学家皮尔逊于20世纪提出的一种计算直线相关的方法

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

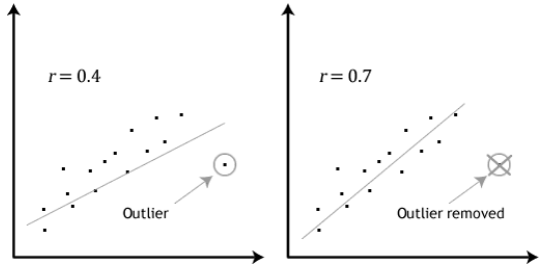
**Assumptions:**

1. For the Pearson **r** correlation, both variables should be normally distributed. 两个变量的总体均符合正态分布
2. There should be no significant outliers. 不能有显著的异常值
3. Each variable should be continuous. 变量是连续变量
4. The two variables have a linear relationship. 两个变量间是线性关系
5. The observations are paired observations. 观测值是成对的观测值
6. Homoscedasticity (assumes that data is equally distributed about the regression line). 同方差性，假设数据在回归线上是均匀分布

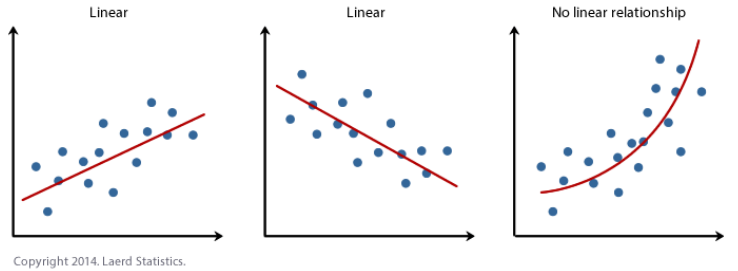
（1）



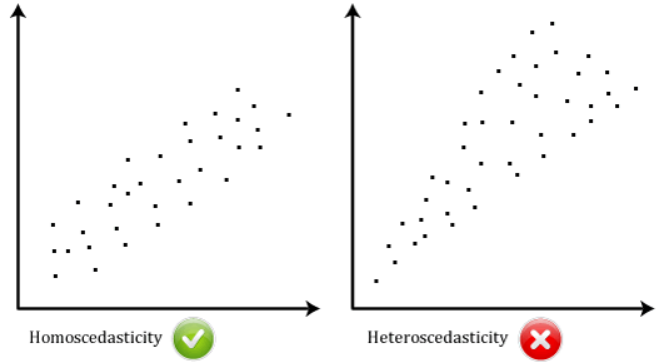

（2）




（4）




（6）



# 相关性（系数）计算

| Infant ID # | Gestational Age (weeks) | $(X-\bar{X})$ | $(X-\bar{X})^2$ |
|---|---|---|---|
| 1 | 34.7 | -3.7 | 13.69 |
| 2 | 36.0 | -2.4 | 5.76 |
| 3 | 29.3 | -9.1 | 82,81 |
| 4 | 40.1 | 1.7 | 2.89 |
| 5 | 35.7 | -2.7 | 7.29 |
| 6 | 42.4 | 4.0 | 16.0 |
| 7 | 40.3 | 1.9 | 3.61 |
| 8 | 37.3 | -1.1 | 1.21 |
| 9 | 40.9 | 2.5 | 6.25 |
| 10 | 38.3 | -0.1 | 0.01 |
| 11 | 38.5 | 0.1 | 0.01 |
| 12 | 41.4 | 3.0 | 9.0 |
| 13 | 39.7 | 1.3 | 1.69 |
| 14 | 39.7 | 1.3 | 1.69 |
| 15 | 41.1 | 2.7 | 7.29 |
| 16 | 38.0 | -0.4 | 0.16 |
| 17 | 38.7 | 0.3 | 0.09 |
| | $\Sigma X = 652.1$ | | $\Sigma(X-\bar{X})^2 = 159.45$ |

| Infant ID# | Birth Weight | $(Y-\bar{Y})$ | $(Y-\bar{Y})^2$ |
|---|---|---|---|
| 1 | 1895 | -1007 | 1,014,049 |
| 2 | 2030 | -872 | 760,384 |
| 3 | 1440 | -1462 | 2,137,444 |
| 4 | 2835 | -67 | 4,489 |
| 5 | 3090 | 188 | 35,344 |
| 6 | 3827 | 925 | 855,625 |
| 7 | 3260 | 358 | 128,164 |
| 8 | 2690 | -212 | 44,944 |
| 9 | 3285 | 383 | 146,689 |
| 10 | 2920 | 18 | 324 |
| 11 | 3430 | 528 | 278,764 |
| 12 | 3657 | 755 | 570,025 |
| 13 | 3685 | 783 | 613,089 |
| 14 | 3345 | 443 | 196,249 |
| 15 | 3260 | 358 | 128,164 |
| 16 | 2680 | -222 | 49,284 |
| 17 | 2005 | -897 | 804,609 |
| | $\Sigma Y = 49,334$ | | $\Sigma(Y-\bar{Y})^2 = 7,767,660$ |

| Infant ID# | $(X-\bar{X})$ | $(Y-\bar{Y})$ | $(X-\bar{X})((Y-\bar{Y})$ |
|---|---|---|---|
| 1 | -3.7 | -1007 | 3725.9 |
| 2 | -2.4 | -872 | 2092.8 |
| 3 | -9,1 | -1462 | 13,304.2 |
| 4 | 1.7 | -67 | -113.9 |
| 5 | -2.7 | 188 | -507.6 |
| 6 | 4.0 | 925 | 3700.0 |
| 7 | 1.9 | 358 | 680.2 |
| 8 | -1.1 | -212 | 233.2 |
| 9 | 2.5 | 383 | 957.5 |
| 10 | -0.1 | 18 | -1.8 |
| 11 | 0.1 | 528 | 52.8 |
| 12 | 3.0 | 755 | 2265.0 |
| 13 | 1.3 | 783 | 1017.9 |
| 14 | 1.3 | 443 | 575.9 |
| 15 | 2.7 | 358 | 966.6 |
| 16 | -0.4 | -222 | 88.8 |
| 17 | 0.3 | -897 | -269.1 |
| | | | Total = 28,768.4 |

$$s_x^2 = \frac{\Sigma(X-\bar{X})^2}{n-1} = \frac{159.45}{16} = 10.0$$

$$s_y^2 = \frac{\Sigma(Y-\bar{Y})^2}{n-1} = \frac{7,767,660}{16} = 485,478.8$$

https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Correlation-Regression/BS704_Correlation-Regression3.html

The covariance of gestational age and birth weight is:

$$Cov(x,y) = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{n-1} = \frac{28,768.4}{16} = 1798.0$$

Finally, we can ow compute the sample correlation coefficient:

$$r = \frac{Cov(x,y)}{\sqrt{\left(s_x^2\right)\left(s_y^2\right)}} = \frac{1798.0}{\sqrt{(10.0)(485,578.8)}} = \frac{1798.0}{2199.4} = 0.82$$

# 2.1.1 Python

numpy.corrcoef

**numpy.corrcoef**(*x*, *y=None*, *rowvar=1*, *bias=0*, *ddof=None*)¶ [source]

Return correlation coefficients.

Please refer to the documentation for cov for more detail. The relationship between the correlation coefficient matrix, *P*, and the covariance matrix, *C*, is

$$P_{ij} = \frac{C_{ij}}{\sqrt{C_{ii} * C_{jj}}}$$

The values of *P* are between -1 and 1, inclusive.

**Parameters:**

**x** : *array_like*
A 1-D or 2-D array containing multiple variables and observations. Each row of *m* represents a variable, and each column a single observation of all those variables. Also see *rowvar* below.

**y** : *array_like, optional*
An additional set of variables and observations. *y* has the same shape as *m*.

**rowvar** : *int, optional*
If *rowvar* is non-zero (default), then each row represents a variable, with observations in the columns. Otherwise, the relationship is transposed: each column represents a variable, while the rows contain observations.

**bias** : *int, optional*
Default normalization is by `(N - 1)`, where `N` is the number of observations (unbiased estimate). If *bias* is 1, then normalization is by `N`. These values can be overridden by using the keyword `ddof` in numpy versions >= 1.5.

**ddof** : *{None, int}, optional*
*New in version 1.5.*
If not `None` normalization is by `(N - ddof)`, where `N` is the number of observations; this overrides the value implied by `bias`. The default value is `None`.

**Returns:**

**out** : *ndarray*
The correlation coefficient matrix of the variables.

# 2.1.1 Python

scipy.stats.pearsonr

## scipy.stats.pearsonr

**scipy.stats.pearsonr(*x*, *y*)** [source]

Calculates a Pearson correlation coefficient and the p-value for testing non-correlation.

The Pearson correlation coefficient measures the linear relationship between two datasets. Strictly speaking, Pearson's correlation requires that each dataset be normally distributed. Like other correlation coefficients, this one varies between -1 and +1 with 0 implying no correlation. Correlations of -1 or +1 imply an exact linear relationship. Positive correlations imply that as x increases, so does y. Negative correlations imply that as x increases, y decreases.

The p-value roughly indicates the probability of an uncorrelated system producing datasets that have a Pearson correlation at least as extreme as the one computed from these datasets. The p-values are not entirely reliable but are probably reasonable for datasets larger than 500 or so.

**Parameters:** **x** : *(N,) array_like*
Input
**y** : *(N,) array_like*
Input

**Returns:** *(Pearson's correlation coefficient,*
2-tailed p-value)

# 2.1.2 How to report Pearson's r in publications

## How to Report Pearson's *r* (Pearson's Correlation Coefficient) in APA Style

The APA has precise requirements for reporting the results of statistical tests, which means as well as getting the basic format right, you need to pay attention to the placing of brackets, punctuation, italics, and so on.

Happily, the basic format for citing Pearson's *r* is not too complex, as you can see here (the color red means you substitute in the appropriate value from your study).

*r*(degress of freedom) = the *r* statistic, *p* = *p* value.

*a)* *r* statistic should be reported at 2 decimal places
b) Remember to drop the leading 0 from both r and p values (i.e. not 0.56, but rather .56)
c) You need not provide the formula for *r*
d) Degree of freedom for r is N-2 (the number of total data points minus 2)

Are there guidelines to interpreting Pearson's correlation coefficient?

Yes, the following guidelines have been proposed:

| Strength of Association | Coefficient, *r* | |
|---|---|---|
| | Positive | Negative |
| Small | .1 to .3 | -0.1 to -0.3 |
| Medium | .3 to .5 | -0.3 to -0.5 |
| Large | .5 to 1.0 | -0.5 to -1.0 |

# 2.1.3 相关系数的假设检验（P-value of Pearson's r）

**相关系数的假设检验:**

❖ 一种是对假设“相关系数 ρ=0” 的 **t-检验（one correlation test）**

零假设H0：总体相关系数与零没有显著差异。在人群中，x和y之间没有显著的线性关系（相关性）。

替代假设Ha：总体相关系数不同于零。人群中x和y之间存在显著的线性关系（相关性）。

较小的p值（<0.05）表示零假设是错误的。可以得出结论，相关系数不同于零，并且存在线性关系。如果p值小于0.05，通常会拒绝零假设。

例：一个实验的两变量之间是否相关?

**检验假设　$H_0$：ρ=0，$H_1$：ρ≠0，α=0.05**

- 在$H_0$为真的情况下，来自(ρ=0)的总体的所有样本相关系数呈对称正态分布，故r的显著性可用t-检验来进行
- 皮尔逊相关系数可以构建一个统计量t
- 是个符合自由度为(n-2)的 t-分布。这里我们就可以使用t-分布进行相关性的检验

$$t = r \cdot \sqrt{\frac{n-2}{1-r^2}}$$

n: # of data points

**关于t检验**(检验r是否显著，即检验r是否不等于零)

| | |
|---|---|
| 1 | 根据r和n计算得到t |
| 2 | 查表得到 在 显著性水平α和自由度(n-2)下，t分布的上α分位点$t_{\alpha/2}$ |
| 3 | 判断$t>t_{\alpha/2}$是否成立，若成立，则r是显著的 |

# 再举个例子

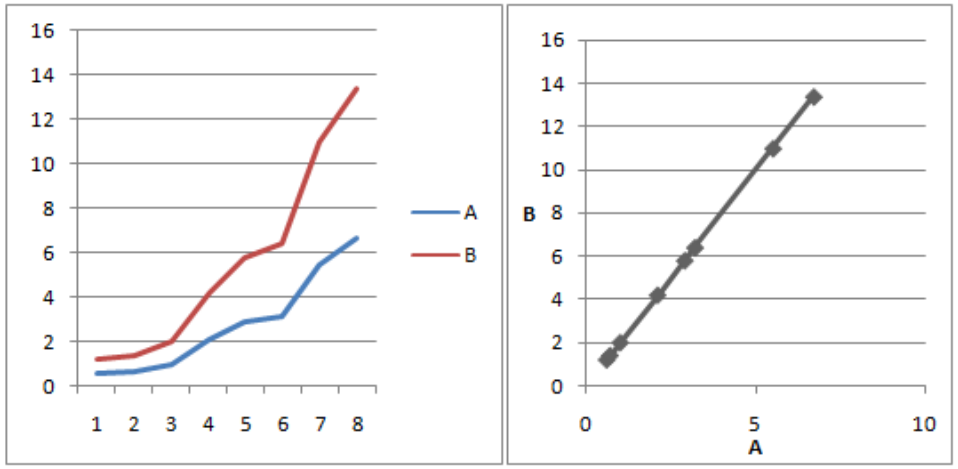
**从皮尔森相关系数定义来看，如果两个基因的表达量呈线性关系（数学上，线性相关指的是直线相关，指数、幂函数、正弦函数等曲线相关不属于线性相关），那么两个基因表达量的就有显著的皮尔森相关系性。下面用几组模拟数值来测试一下：**

测试1：两个基因A、B，他们在8个样本中的表达量值如下：

表 1　基因 A、B 在 8 个样本中的表达量值

| 样本编号 | 样本 1 | 样本 2 | 样本 3 | 样本 4 | 样本 5 | 样本 6 | 样本 7 | 样本 8 |
|---|---|---|---|---|---|---|---|---|
| A | 0.6 | 0.7 | 1 | 2.1 | 2.9 | 3.2 | 5.5 | 6.7 |
| B | 1.2 | 1.4 | 2 | 4.2 | 5.8 | 6.4 | 11 | 13.4 |

图 1　基因 A、B 在 8 个样本中的表达量示意图



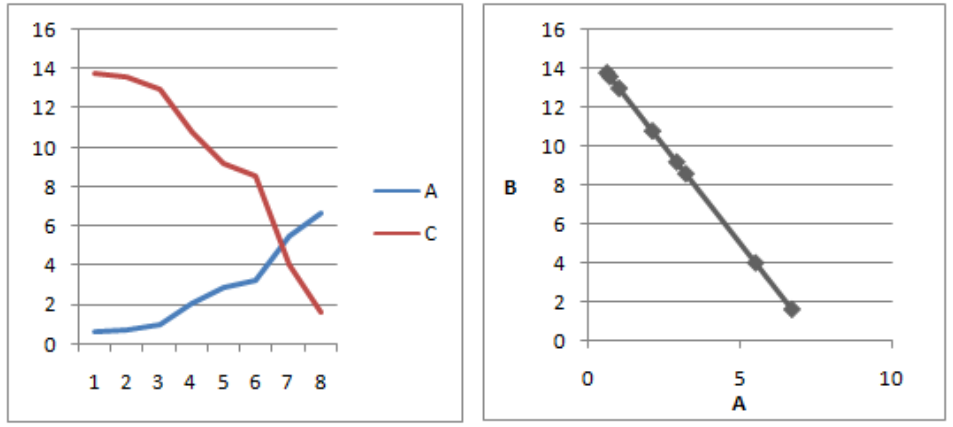

计算得出，他们的皮尔森相关系数 $r$=1，P-vlaue≈0。

测试2：两个基因A、C，他们的关系是C=15-2A，在8个样本中的表达量值如下：

表 2　基因 A、C 在 8 个样本中的表达量值

| 样本编号 | 样本 1 | 样本 2 | 样本 3 | 样本 4 | 样本 5 | 样本 6 | 样本 7 | 样本 8 |
|---|---|---|---|---|---|---|---|---|
| A | 0.6 | 0.7 | 1 | 2.1 | 2.9 | 3.2 | 5.5 | 6.7 |
| C | 13.8 | 13.6 | 13 | 10.8 | 9.2 | 8.6 | 4 | 1.6 |

图 2 基因 A、C 在 8 个样本中的表达量示意图




计算得出，他们的皮尔森相关系数 r=-1，P-vlaue≈0。
从以上可以直观看出，如果两个基因的表达量呈线性关系，则具有显著的皮尔森相关性。如果两个基因“共舞”（如图1），则两者正相关；如果“你要往东，我偏往西”（如图2），则两者负相关。

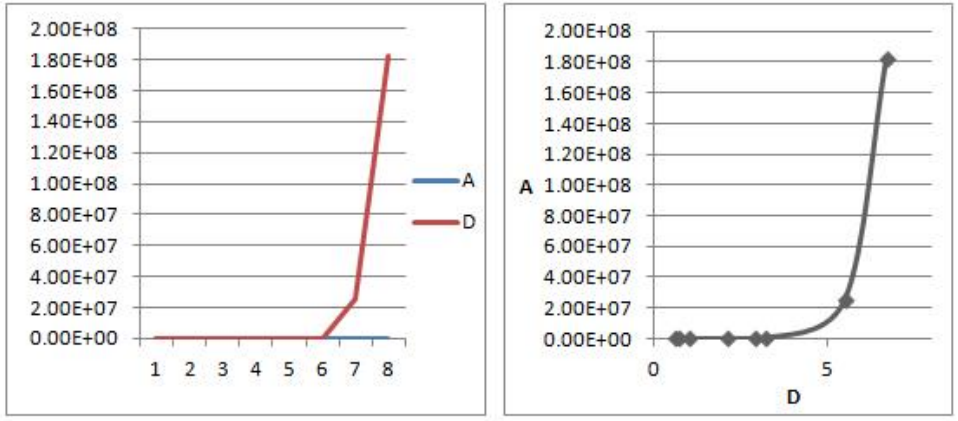
**以上是两个基因呈线性关系的结果。如果两者呈非线性关系，例如幂函数关系（曲线关系），那又如何呢？ 我们再试试：**

测试3：两个基因A、D，他们的关系是D=A10，在8个样本中的表达量值如下：

表 3　基因 A、D 在 8 个样本中的表达量值

| 样本编号 | 样本 1 | 样本 2 | 样本 3 | 样本 4 | 样本 5 | 样本 6 | 样本 7 | 样本 8 |
|---|---|---|---|---|---|---|---|---|
| A | 0.6 | 0.7 | 1 | 2.1 | 2.9 | 3.2 | 5.5 | 6.7 |
| D | 6.0E-3 | 2.8E-2 | 1 | 1.7E3 | 4.2E4 | 1.1E5 | 2.5E7 | 1.8E8 |

图 3 基因 A、D 在 8 个样本中的表达量示意图




计算得出，他们的皮尔森相关系数等于 0.77，P value= 0.0267

可以看到，基因A、D相关系数，无论数值还是显著性都下降了。皮尔森相关系数是一种线性相关系数，因此如果两个变量呈线性关系的时候，具有最大的显著性。对于非线性关系（例如A、D的幂函数关系），则其对相关性的检测功效会下降。但在生物体内的许多调控关系，例如转录因子与靶基因、小干扰RNA与靶基因，可能都是非线性关系，那么是否有更合适的相关系数检测方法呢？

其实可以考虑另外一个相关系数计算方法：斯皮尔曼等级相关。

# 2.2 三大相关系数之- 斯皮尔曼秩相关性系数（Spearman's Rank）

非线性关系

Spearman斯皮尔曼相关性系数，通常也叫斯皮尔曼秩相关系数。“秩”，可以理解成就是一种顺序或者排序，它就是**根据原始数据的排序位置进行求解**，所以又称为“等级差数法”。对原始变量的分布不作要求，属于非参数统计方法，适用范围要广些。 相关系数 ρ

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

实际应用中， 变量间的连结是无关紧要的， 于是可以通过简单的步骤计算 ρ.
被观测的两个变量的等级的差值 $d_i=(x_i-y_i)$ ，则 ρ 为

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}.$$

d 为两列成对变量的等级差数

| 变量 $X_i$ | 降序排名 | $x_i$ 的排名 |
|---|---|---|
| 0.8 | 5 | 5 |
| 1.2 | 4 | $\frac{4+3}{2} = 3.5$ |
| 1.2 | 3 | $\frac{4+3}{2} = 3.5$ |
| 2.3 | 2 | 2 |
| 18 | 1 | 1 |

**Spearman's Rank Correlation Coefficient**

$$\rho = \frac{\sum_{i=1}^{n}(R(x_i) - \overline{R(x)})(R(y_i) - \overline{R(y)})}{\sqrt{\sum_{i=1}^{n}(R(x_i) - \overline{R(x)})^2 \cdot \sum_{i=1}^{n}(R(y_i) - \overline{R(y)})^2}} = 1 - \frac{6\sum_{i=1}^{n}(R(x_i) - R(y_i))^2}{n(n^2-1)}$$

$$\begin{aligned} Where,\ R(x_i) &= rank\ of\ x_i \\ R(y_i) &= rank\ of\ y_i \\ \overline{R(x)} &= mean\ rank\ of\ x \\ \overline{R(y)} &= mean\ rank\ of\ y \\ n &= number\ of\ pairs \end{aligned}$$

**Assumptions:**

- Pairs of observations are independent.
- Two variables should be measured on an ordinal, interval or ratio scale.
- It assumes that there is a monotonic relationship between the two variables.

简单点说，就是无论两个变量的数据如何变化，符合什么样的分布，我们只关心每个数值在变量内的排列顺序。如果两个变量的对应值，在各组内的排序顺位是相同或类似的，则具有显著的相关性。

# 手动计算

- 第1步：创建一个数据表
- 第2步：首先对两个数据集进行排序。数据排名可以通过将排名“1”分配给列中最大的数字，“2”分配给第二大的数字，依此类推。最小的值通常会得到最低的排名。这两组测量都应该进行
- 第3步：向数据集中添加第三列d，这里d表示等级之间的差异。例如，如果第一个学生的物理排名是3，数学排名是5，那么排名的差异是3。在第四列中，将d值平方

| History | Rank | Geography | Rank | d | d square |
|---|---|---|---|---|---|
| 35 | 3 | 30 | 5 | 2 | 4 |
| 23 | 5 | 33 | 3 | 2 | 4 |
| 47 | 1 | 45 | 2 | 1 | 1 |
| 17 | 6 | 23 | 6 | 0 | 0 |
| 10 | 7 | 8 | 8 | 1 | 1 |
| 43 | 2 | 49 | 1 | 1 | 1 |
| 9 | 8 | 12 | 7 | 1 | 1 |
| 6 | 9 | 4 | 9 | 0 | 0 |
| 28 | 4 | 31 | 4 | 0 | 0 |
| | | | | | **12** |

第4步-将所有的d平方值相加，即12*(∑d square)

第5步-在公式中插入这些值

$$r_R = 1 - \frac{6\Sigma_i d_i^2}{n(n^2-1)}$$

=1-(6*12)/(9(81-1))

=1-72/720

=1-01

=0.9

举个例子，例如表3的数值，用斯皮尔曼等级相关计算相关系数，将发生如下变化。

表 3　基因 A、D 在 8 个样本中的表达量值

| 样本编号 | 样本 1 | 样本 2 | 样本 3 | 样本 4 | 样本 5 | 样本 6 | 样本 7 | 样本 8 |
|---|---|---|---|---|---|---|---|---|
| A | 0.6 | 0.7 | 1 | 2.1 | 2.9 | 3.2 | 5.5 | 6.7 |
| D | 6.0E-3 | 2.8E-2 | 1 | 1.7E3 | 4.2E4 | 1.1E5 | 2.5E7 | 1.8E8 |

图 3 基因 A、D 在 8 个样本中的表达量示意图

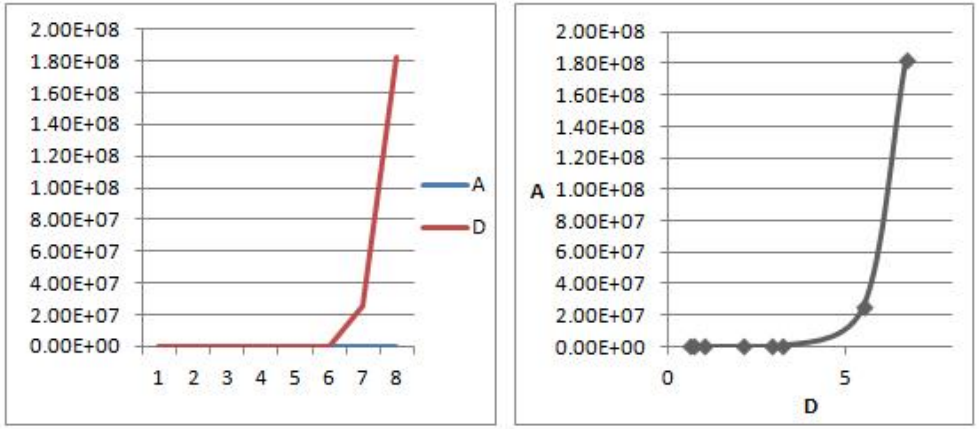


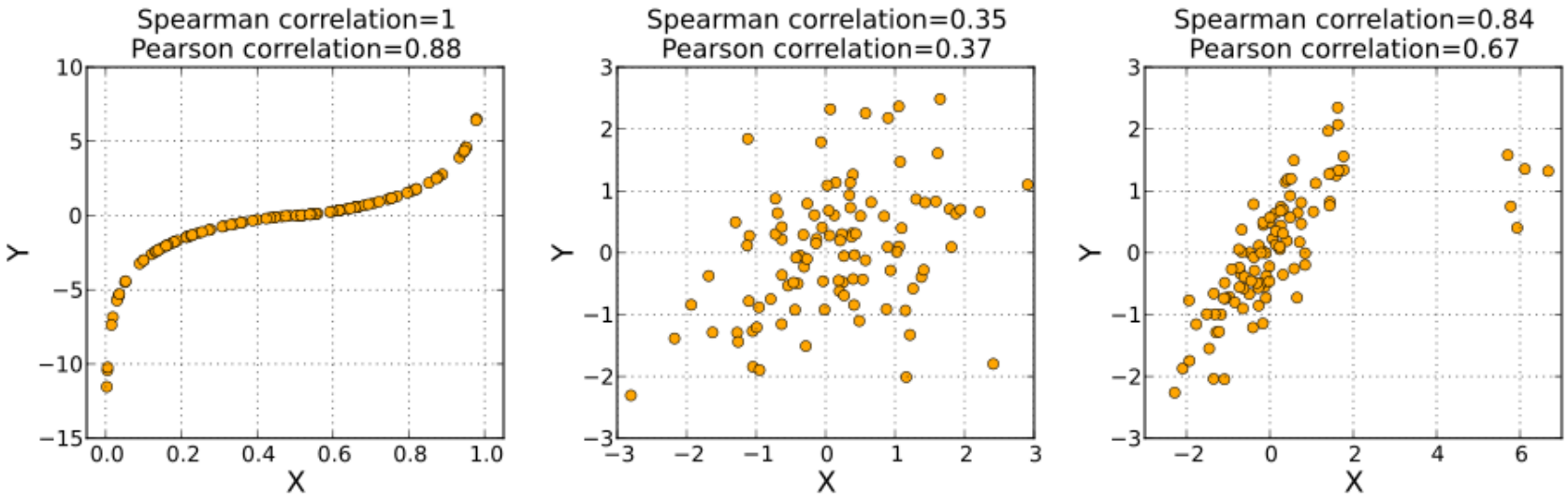

表 4　斯皮尔曼等级排列

| 样本编号 | 样本 1 | 样本 2 | 样本 3 | 样本 4 | 样本 5 | 样本 6 | 样本 7 | 样本 8 |
|---|---|---|---|---|---|---|---|---|
| A 表达量 | 0.6 | 0.7 | 1 | 2.1 | 2.9 | 3.2 | 5.5 | 6.7 |
| A 排序等级 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| D 表达量 | 6.0E-3 | 2.8E-2 | 1 | 1.7E3 | 4.2E4 | 1.1E5 | 2.5E7 | 1.8E8 |
| D 排序等级 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| d(等级差) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

备注：排序等级就是这个数值在组内从小到大排列的序位号。

他们的皮尔森相关系数等于 0.77，P value= 0.0267

利用斯皮尔曼等级相关计算A、D基因表达量的相关性，结果是:
r=1，p-value = 4.96e-05
这里斯皮尔曼等级相关的显著性显然高于皮尔森相关。这是因为虽然两个基因的表达量是非线性关系，但两个基因表达量在所有样本中的排列顺序是完全相同的，因为具有极显著的斯皮尔曼等级相关性



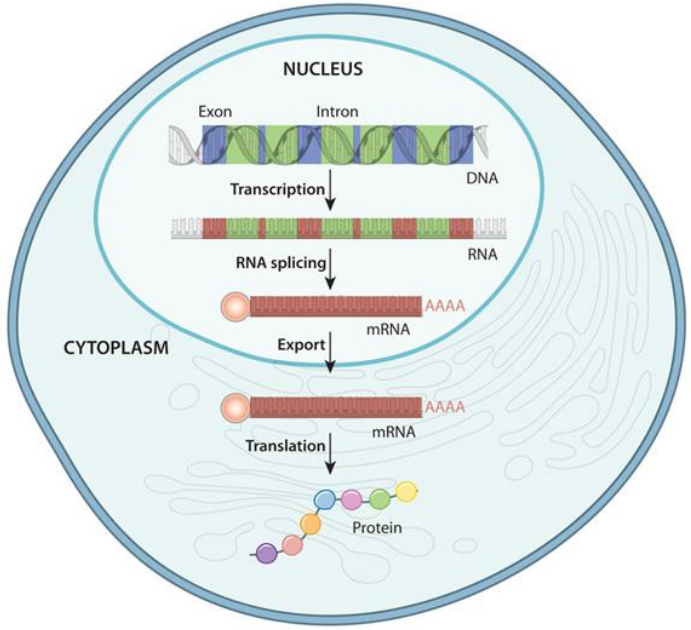

TIPS 适用范围：

斯皮尔曼等级相关系数对数据条件的要求没有皮尔逊相关系数严格，只要两个变量的观测值是成对的等级评定资料，或者是由连续变量观测资料转化得到的等级资料，不论两个变量的总体分布形态、样本容量的大小如何，都可以用斯皮尔曼等级相关系数来进行研究

## 总结

皮尔森相关和斯皮尔曼等级相关，都是在计算基因共表达或多组学贯穿分析时常用的相关性度量方法。因为基因间调控方式可能并非线性，加上实验误差、检测误差等因素的干扰，皮尔森相关的显著性可能会下降。而斯皮尔曼等级相关可能可以弥补以上的缺陷，因此一些软件也提供了这个选择。

但由于生物体调控方式的复杂性，例如多个基因联合调控一个下游基因，我们并不能武断决定哪一种相关性计算方式最佳，还是需要根据具体情况定制个性化的分析策略



# 实践使用注意事项

## 适用范围与计算方法选择

Spearman 和Pearson相关系数在算法上完全相同. 只是Pearson相关系数是用原来的数值计算积差相关系数, 而Spearman是用原来数值的秩次计算积差相关系数。

1、Pearson相关系数适用条件为两个变量间有线性关系、变量是连续变量、变量均符合正态分布。
2、若上述有条件不满足则考虑用Spearman相关系数
3、对于同一量纲数据建议Pearson，例如mRNA基因表达量数据，计算不同mRNA表达量的相关系数；对于不同量纲数据，可考虑Spearman相关系数，例如mRNA表达量与某表型数据（株高、产果量、次生化合物含量等）

## 相关系数的缺点与注意事项

需要指出的是，相关系数有一个明显的缺点，即它接近于1的程度与数据组数n相关，这容易给人一种假象。因为，当n较小时，相关系数的波动较大，对有些样本相关系数的绝对值易接近于1；当n较大时，相关系数的绝对值容易偏小。特别是当n=2时，相关系数的绝对值总为1。因此在样本容量n较小时，我们仅凭相关系数较大就判定变量x与y之间有密切的线性关系是不妥当的。

因此比如高通量测序项目，一般建议**10个以上样本**才计算相关系数，这样其可靠性更高。

# 2.3 三大相关系数之- 肯德尔秩相关性系数（Kendall’s Tau）

非线性关系

前面我们已经讨论了 **Pearson** 相关系数和 **Spearman** 秩相关系数，它们可以检测连续变量间的相关性，并且 **Spearman** 秩相关系数还能够检测有序的离散变量间的相关系数。今天我们再讨论一个能够检测有序变量相关性的系数：**Kendall** 秩相关系数。这里有序变量既包括实数变量，也包括可以排序的类别变量，比如名次、年龄段等

**Kendall** 秩相关系数以 Maurice Kendall 命名的，并经常用希腊字母 **τ（tau）** 表示其值。是一个非参数性质（与分布无关）的秩统计参数，是用来度量两个**有序变量**之间**单调关系**强弱的相关系数，它的取值范围是 [-1,1]，绝对值越大，表示单调相关性越强，取值为 0 时表示完全不相关

# 三大相关系数之- 肯德尔秩相关性系数（Kendall's Tau）

定义：原始的 Kendall 秩相关系数定义在一致对 (concordant pairs) 和分歧对 (discordant pairs) 的概念上。所谓一致对，就是两个变量取值的相对关系一致；分歧对则是指它们的相对关系不一致。**这么说有点难以理解，简单的说就是将两个变量进行排序，判断两者的排序值是否一致。如果一致则为1，如果倒叙则为-1**

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 身高 (cm) | 155 | 160 | 165 | 173 | 178 | 185 | 187 | 190 |
| 体重 (kg) | 110 | 140 | 180 | 160 | 80 | 60 | 50 | 65 |

**如按身高正确排序，则体重是乱序的**

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 身高 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 体重 | 3 | 4 | 1 | 2 | 5 | 7 | 8 | 6 |

| | | | | | | | | | **同序数** |
|---|---|---|---|---|---|---|---|---|---|
| A | 3 | ✓ | | | ✓ | ✓ | ✓ | ✓ | 5 |
| B | 4 | | | | ✓ | ✓ | ✓ | ✓ | 4 |
| C | 1 | | ✓ | ✓ | ✓ | ✓ | ✓ | | 5 |
| D | 2 | | | ✓ | ✓ | ✓ | ✓ | | 4 |
| E | 5 | | | | ✓ | ✓ | ✓ | | 3 |
| F | 7 | | | | | ✓ | | | 1 |
| G | 8 | | | | | | | | 0 |
| H | 6 | | | | | | | | 0 |

肯德尔相关系数与斯皮尔曼相关系数对数据条件的要求相同。用于反映分类变量相关性的指标，适用于两个分类变量均为有序分类的情况

比如评委对选手的评分（优，中，差），我们想看两个（或者多个）评委对即为选手的评价标准是否一致；或者医院的尿糖化验报告，想检验各个医院对尿糖的化验结果是否一致，这时就可以用Kendall相关系数进行衡量。

$$\text{tau} = \frac{C - D}{\frac{1}{2}N(N-1)}$$

C = 5 + 4 + 5 + 4 + 3 + 1 + 0 + 0 = 22

D = 28 – 22 (总对数减去同序对数为异序对数)

tau = ((22-6)/28)=0.57，说明变量之间是正相关的

**Pearson's Correlation Coefficient**

$$r = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum(X - \overline{X})^2 \cdot \sum(Y - \overline{Y})^2}}$$

$$Where,\ \overline{X} = mean\ of\ X\ variable$$
$$\overline{Y} = mean\ of\ Y\ variable$$

**Assumptions:**

- Each observation should have a pair of values.
- Each variable should be continuous.
- Each variable should be normally distributed.
- It should be the absence of outliers.
- It assumes linearity and homoscedasticity.

**Spearman's Rank Correlation Coefficient**

$$\rho = \frac{\sum_{i=1}^{n}(R(x_i) - \overline{R(x)})(R(y_i) - \overline{R(y)})}{\sqrt{\sum_{i=1}^{n}(R(x_i) - \overline{R(x)})^2 \cdot \sum_{i=1}^{n}(R(y_i) - \overline{R(y)})^2}} = 1 - \frac{6\sum_{i=1}^{n}(R(x_i) - R(y_i))^2}{n(n^2 - 1)}$$

$$Where,\ R(x_i) = rank\ of\ x_i$$
$$R(y_i) = rank\ of\ y_i$$
$$\overline{R(x)} = mean\ rank\ of\ x$$
$$\overline{R(y)} = mean\ rank\ of\ y$$
$$n = number\ of\ pairs$$

**Assumptions:**

- Pairs of observations are independent.
- Two variables should be measured on an ordinal, interval or ratio scale.
- It assumes that there is a monotonic relationship between the two variables.

**Kendall's Tau Coefficient**

$$\tau = \frac{n_c - n_d}{n_c + n_d} = \frac{n_c - n_d}{n(n-1)/2}$$

$$Where,\ n_c = number\ of\ concordant\ pairs$$
$$n_d = number\ of\ discordant\ pairs$$
$$n = number\ of\ pairs$$

**Assumptions:**

- It's the same as assumptions of *Spearman's rank correlation coefficient*

附示例的 python 代码

```
>>> from scipy.stats import kendalltau
>>> import numpy as np
>>> x=[1,2,3,4,5,6,7,8,9,10]
>>> y=[1,5,2,4,3,7,6,8,9,10]
>>> kendalltau(x,y)
(0.7777777777777779, 0.0017451191944018172)
>>> x=[1,1,1,2,2,2,2,3,3,4]
>>> y=[1,1,1,1,1,1,2,2,2,2]
>>> kendalltau(x,y)
(0.72456883730947197, 0.0035417200011750309)
```

其中，`kendalltau` 返回的第二个结果是 p-value，其具体含义可参考官方文档。

# Correlation: Pearson r, Spearman's ρ, Kendal's Tau τ,

| Pearson's r | *Spearman's ρ* | Kendall's tau |
|---|---|---|
| Ratio/interval data | Ordinal | Ordinal |
| | | Better interpretation than *ρ* |
| | Monotonicity measure | Monotonicity measure |
| | Confidence intervals for Spearman's $r_S$ are less reliable and less interpretable than confidence intervals for Kendall's τ-parameters | τ is also more tractable mathematically, particularly when ties are present. Also it is preferred when there are small samples (~20) or some outliers |


- 肯德尔相关系数与斯皮尔曼相关系数对数据条件的要求相同。
- Kendall相关比Spearman相关更为稳定且有效。这意味着当存在小样本或一些异常值时，肯德尔相关是首选的。
- Spearman rho值往往高于Kendall tau值。

Non-parametric correlations are less powerful because they use less information in their calculations. In the case of *Pearson's correlation* uses information about the mean and deviation from the mean, while non-parametric correlations use only the ordinal information and scores of pairs

# More …

- A ranking of countries based on the highest total number COVID-19 cases was calculated on 16 June 2020 based on data from the open-source ‘World-O-Meter’ online data repository.
- The FIFA ranking is a point-based system derived by adding points a national football team gains from playing international matches over the period of the last 4 years.

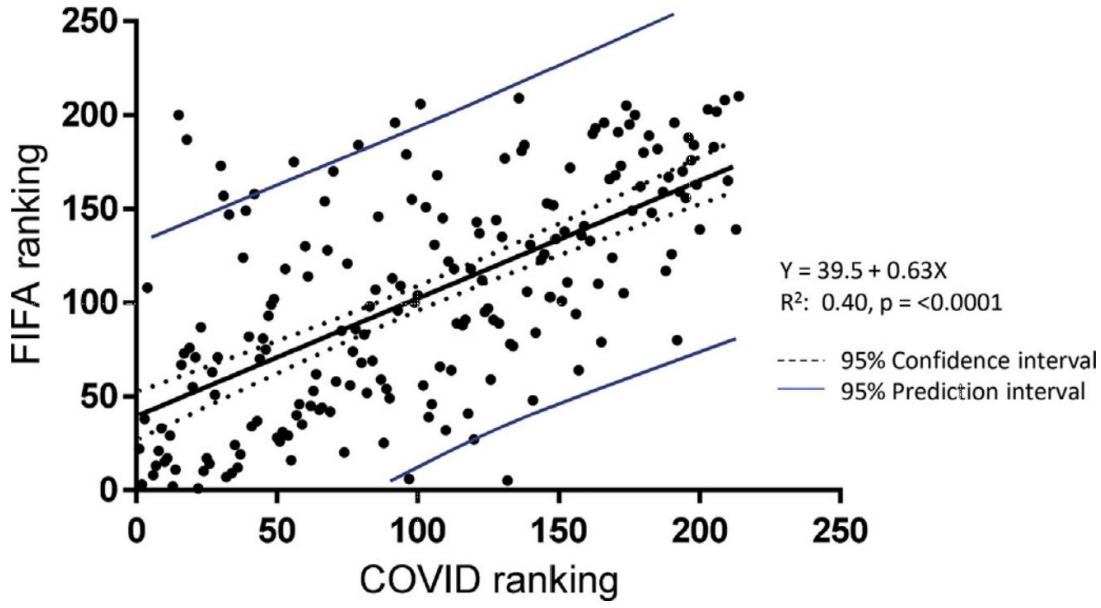



**Fig. 1.** Scatterplot showing the association between national football team FIFA ranking and the respective country's COVID-19 ranking based on total confirmed COVID-19 cases. There was a strong correlation between the two rankings ($R^2$ 0.40, $p < 0.0001$).

Clinical Microbiology and Infection 27 (2021) 291–292

Contents lists available at ScienceDirect

Clinical Microbiology and Infection

journal homepage: www.clinicalmicrobiologyandinfection.com

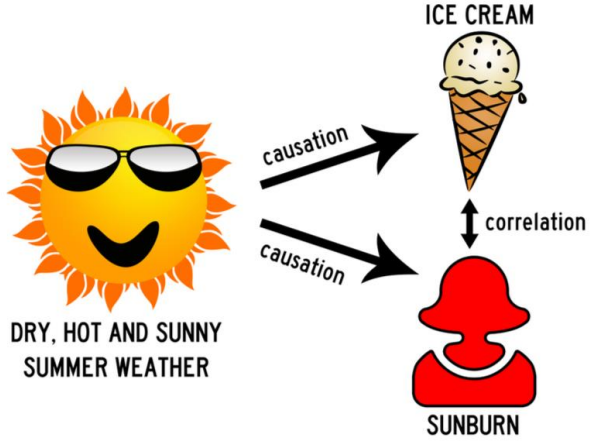
Letter to the Editor

## Football and COVID-19 risk: correlation is not causation

Fares Ayoub [1], Toshiro Sato [2, 3], Atsushi Sakuraba [1, *]

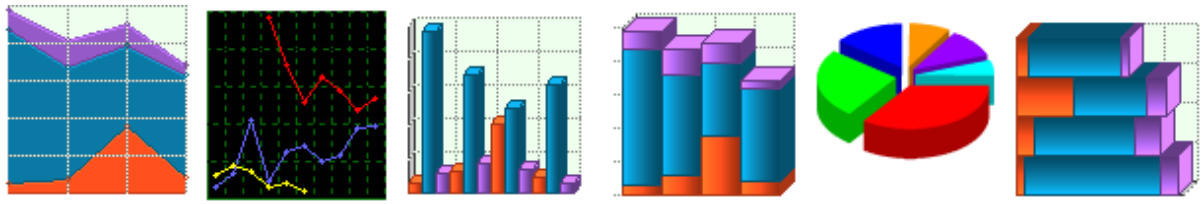
[1] Section of Gastroenterology, Hepatology, and Nutrition, University of Chicago Medicine, Chicago, IL, USA
[2] Department of Organoid Medicine, Keio University School of Medicine, Tokyo, Japan
[3] Corona Virus Task Force, Keio University School of Medicine, Tokyo, Japan

While much of the published COVID-19 observational research has been fuelled by genuine scientific curiosity, we must not let our desperate effort to identify at-risk populations and effective treatments erase our appreciation of the well-established hierarchy of evidence. Instead, we should not forget that outcomes of observational studies are only hypothesis forming, allowing closer examination of the situation to identify whether a true causation can be established between the two variables.





Letter to the Editor / Clinical Microbiology and Infection 27 (2021) 291-292




# Unit 4: Covariance, Correlation & Regression

# 3. Regression（回归）

- **What is Regression Analysis?**

用于预测、时间序列建模和发现变量之间的**因果关系。**例如，通过回归分析，能较好地研究驾驶员鲁莽驾驶与交通事故次数之间的关系

  - Regression analysis is a form of predictive modelling technique which investigates the relationship between a **dependent** (target) and **independent variable (s)** (predictor)

    一种研究自变量和因变量之间关系的预测模型

- **Why do we use Regression Analysis?**
  - It indicates the **significant relationships** between dependent variable and independent variable.
  - It indicates the **strength of impact** of multiple independent variables on a dependent variable

- **What are the types of Regressions?**
  - **Linear Regression** 线性回归
  - **Logistic Regression** 逻辑回归
  - Polynomial Regression 多项式回归
  - Stepwise Regression 逐步回归
  - Ridge Regression 岭回归
  - Lasso Regression 套索回归
  - ElasticNet Regression 弹性网回归

线性回归只能用于表现出线性或近似线性关系的数据

https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/

# 3.1 Linear regression description

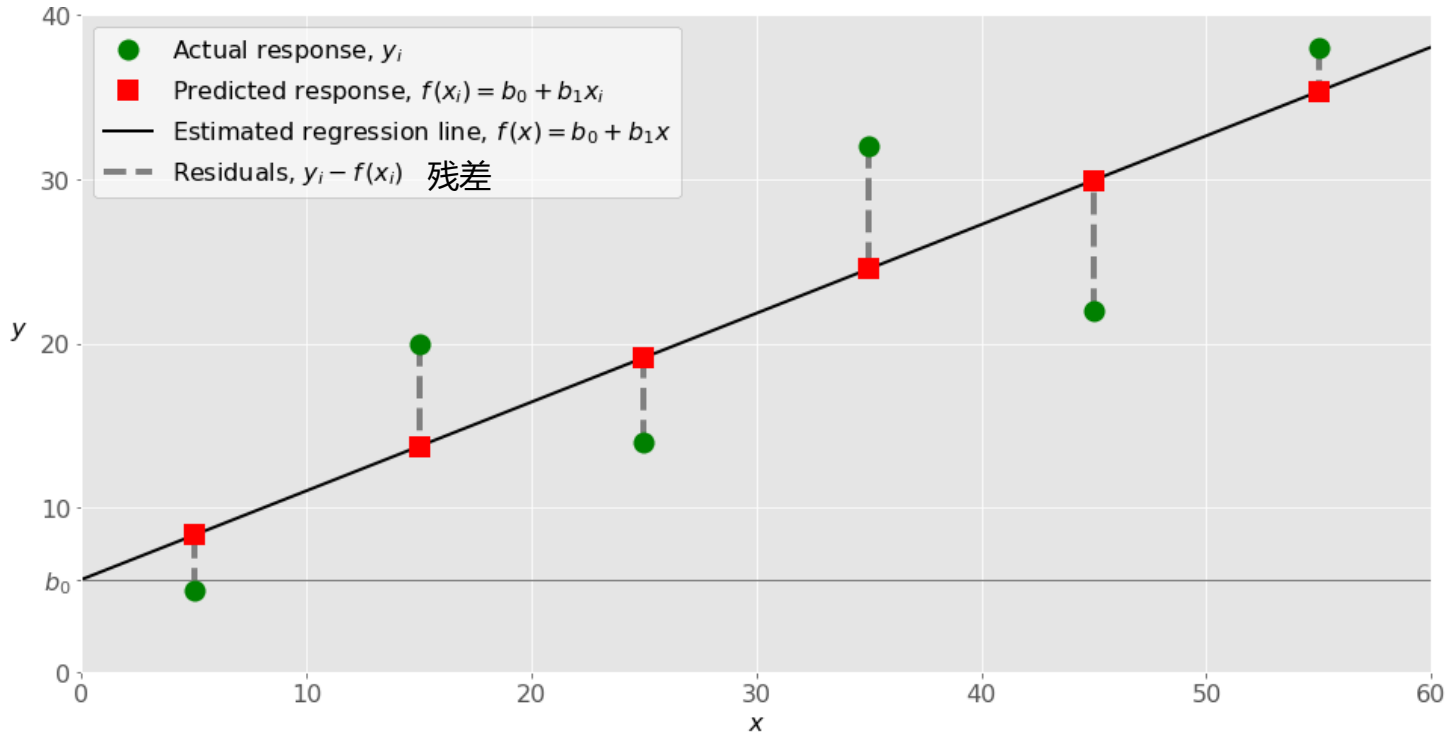
- **Linear regression**: $Y = a + bX + e$ 简单线性回归用于估计因变量y和单个预测因子x之间的线性关系 简单线性回归
- **Multiple linear regression**: $Y = a + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_tX_t + e$ 多元线性回归

"a functional relationship between two or more correlated variables that is often empirically determined from data and is used especially to predict values of one variable when given values of the others"

Where:

- Y = the variable that you are trying to predict (dependent variable). 想预测的变量（因变量）
- X = the variable that you are using to predict Y (independent variable). 用于预测Y的变量（自变量）
- a = the intercept. 截距
- b = the slope. 斜率
- e = the regression residual. 回归残差

# 3.2 Assessing the regression model

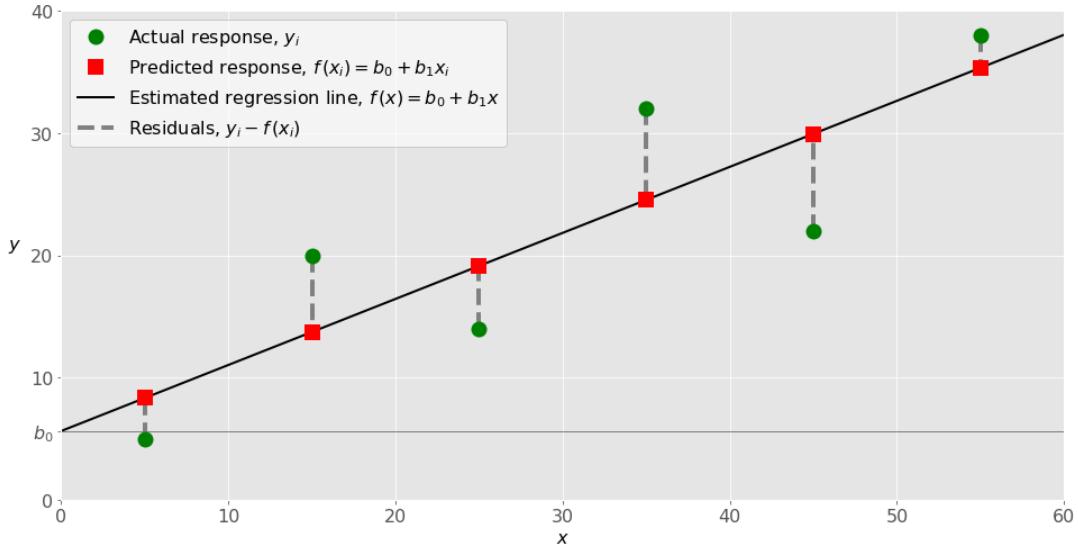


# 3.2.1 Use $R^2$ to measure the Goodness-of-fit of the regression model 决定系数/ 拟合度

## 如何判断 回归方程 的拟合程度?

Q：在对数据进行线性回归计算之后，我们能够得出相应函数的系数，那么我们如何知道得出的这个系数对函数结果的影响有多强呢?

A：用 **coefficient of determination (决定系数)** 来度量因变量的变异中可由自变量解释部分所占的比例，来判断 回归方程 拟合的程度，以此来判断统计模型的解释力。

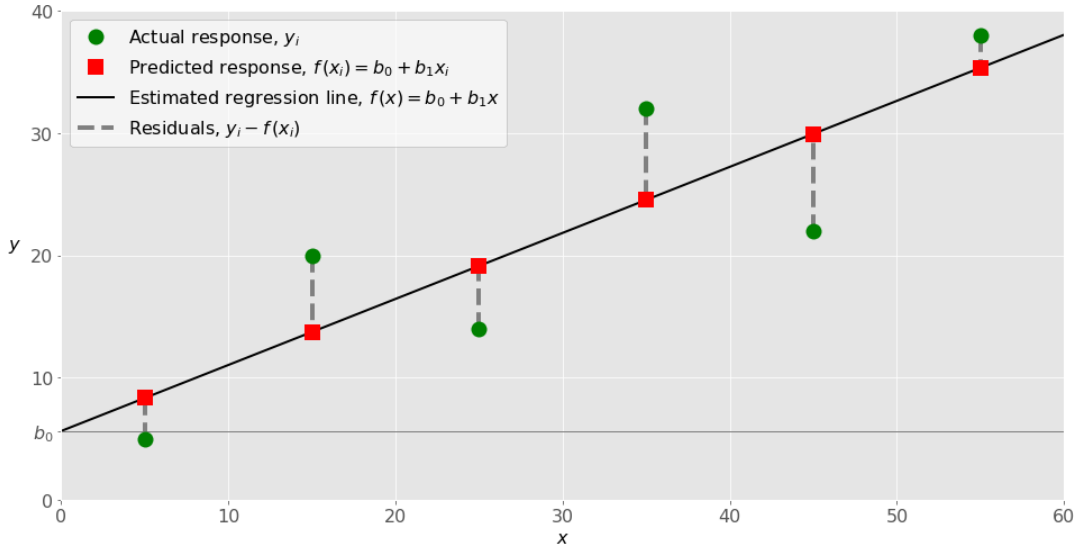


# 3.2.1 Use $R^2$ to measure the Goodness-of-fit of the regression model
# 决定系数/ 拟合度

## 如何判断 回归方程 的拟合程度?

Q：在对数据进行线性回归计算之后，我们能够得出相应函数的系数，那么我们如何知道得出的这个系数对函数结果的影响有多强呢?

A：用 **coefficient of determination (决定系数)** 来度量因变量的变异中可由自变量解释部分所占的比例，来判断 回归方程 拟合的程度，以此来判断统计模型的解释力。




R-Squared ($R^2$ or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. **In other words, r-squared tells how well the data fit the regression model (the goodness of fit).**

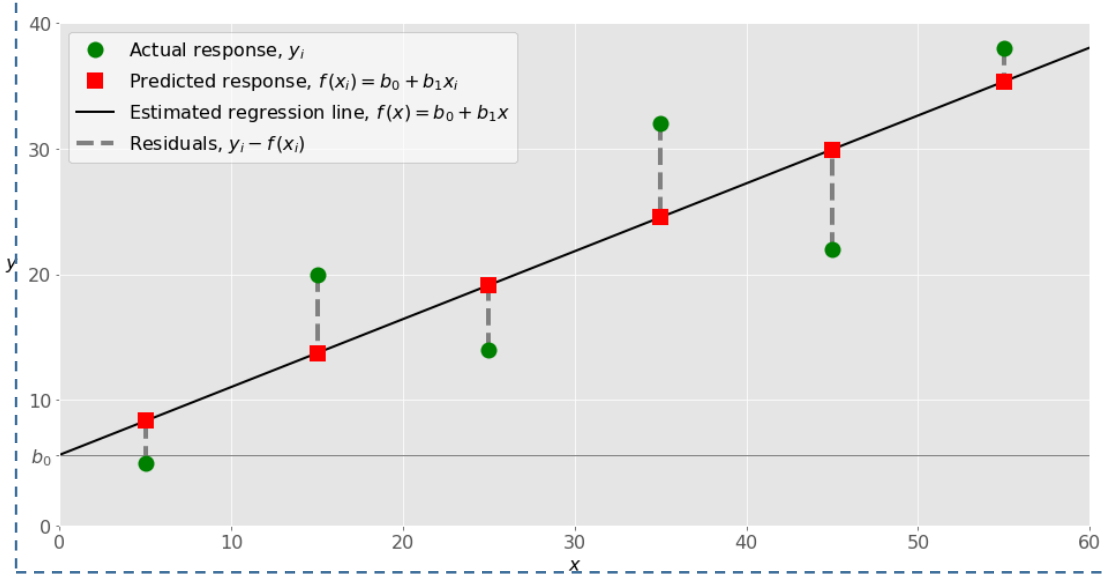
$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\text{Variance explained by the model}}{\text{Total variance}}$$

R平方（$R^2$或决定系数）是回归模型中的一种统计度量，用于确定因变量中可由自变量解释的方差比例。
换句话说，R平方表示数据与回归模型的拟合程度（拟合优度）

决定系数的大小决定了相关的密切程度。当$R^2$ 越接近1时，表示相关的函数参考价值越高；相反，越接近0时，表示参考价值越低。

# 3.2.2 Assess the regression: R-Squared ($R^2$)

A




B

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

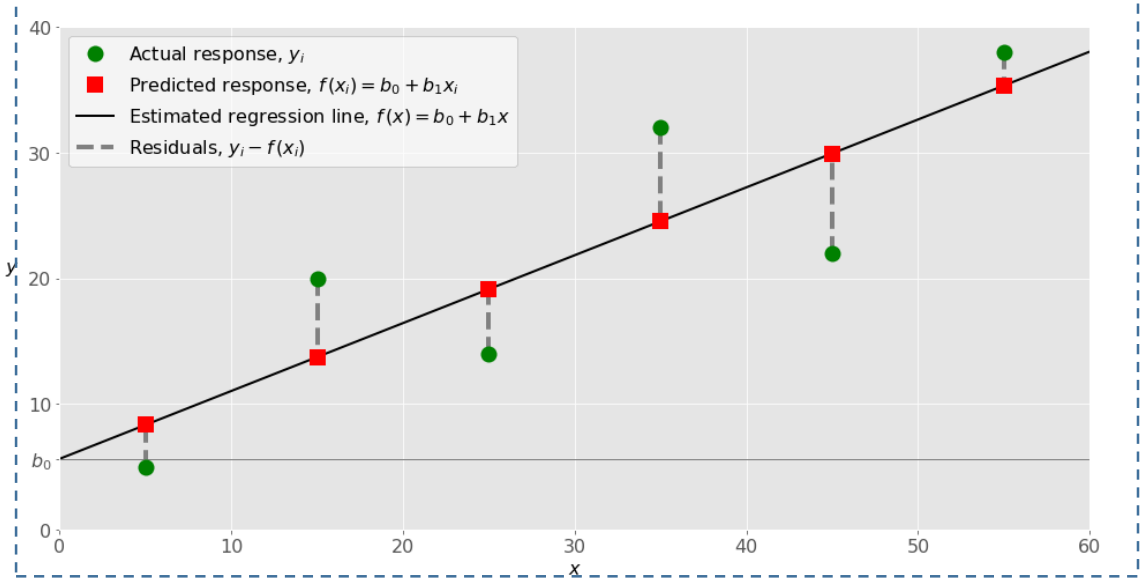
$y_i$ ：观察值（real value）

$\bar{y}$ ：观察值平均值（mean of real value）

$\hat{y}_i$ ：模型预测值（predicted value from model）

# 3.2.2 Assess the regression: R-Squared ($R^2$)

A




B

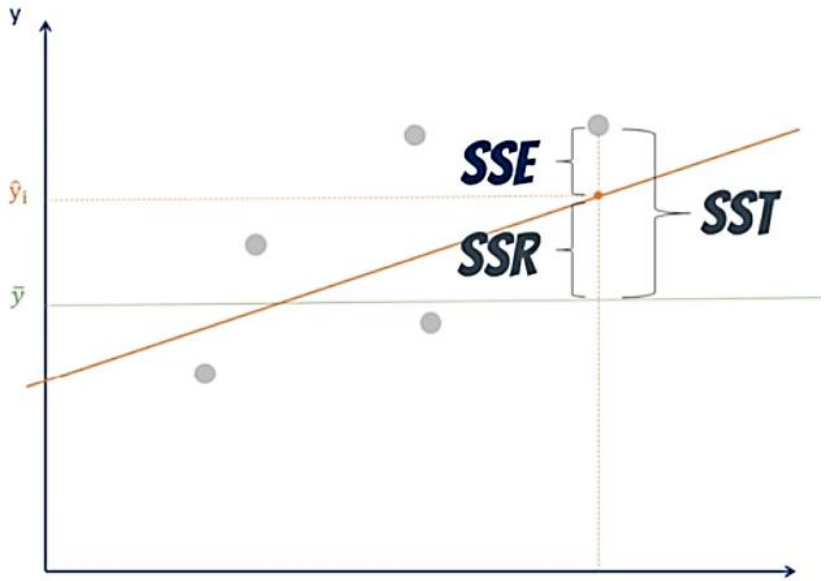
$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}e_i^2$$

$y_i$ ：观察值（real value）

$\bar{y}$ ：观察值平均值（mean of real value）

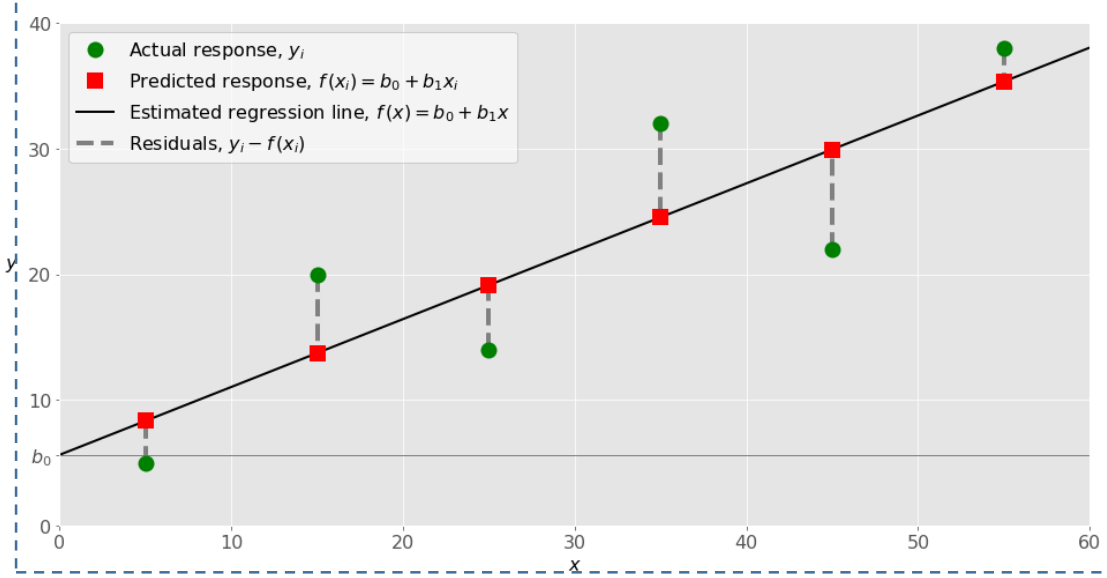
$\hat{y}_i$ ：模型预测值（predicted value from model）

C



# 3.2.2 Assess the regression: R-Squared ($R^2$)

A




B

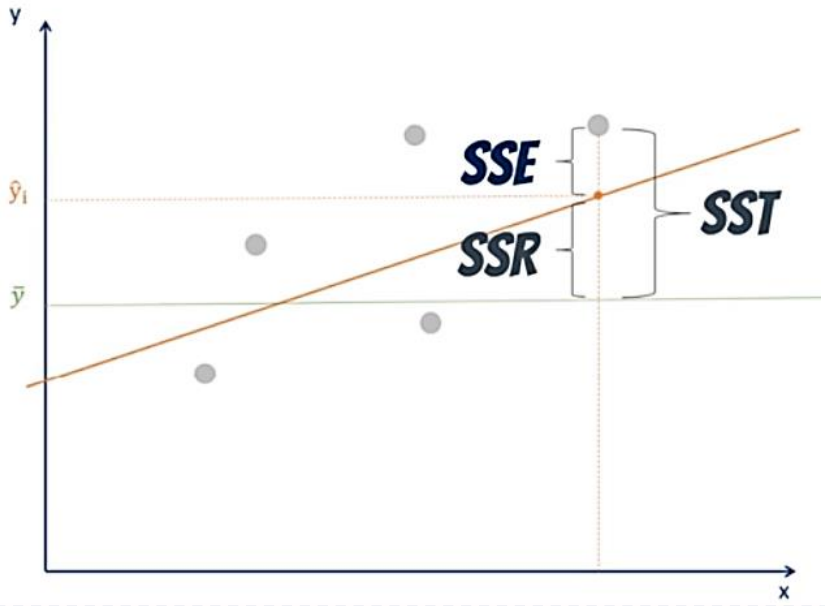
$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}e_i^2$$

$y_i$ ：观察值（real value）

$\bar{y}$ ：观察值平均值（mean of real value）

$\hat{y}_i$ ：模型预测值（predicted value from model）

C




D

**SST**
**Or TSS**

Sum of squares total

平方和
(数据的总变异)

**SSR**

Sum of squares regression

平方回归和
(能被回归模型解释的变异量)

**SSE**
**Or ESS or RSS**

Sum of squares error/Residual sum of square

平方误差和
(无法解释的变异量)

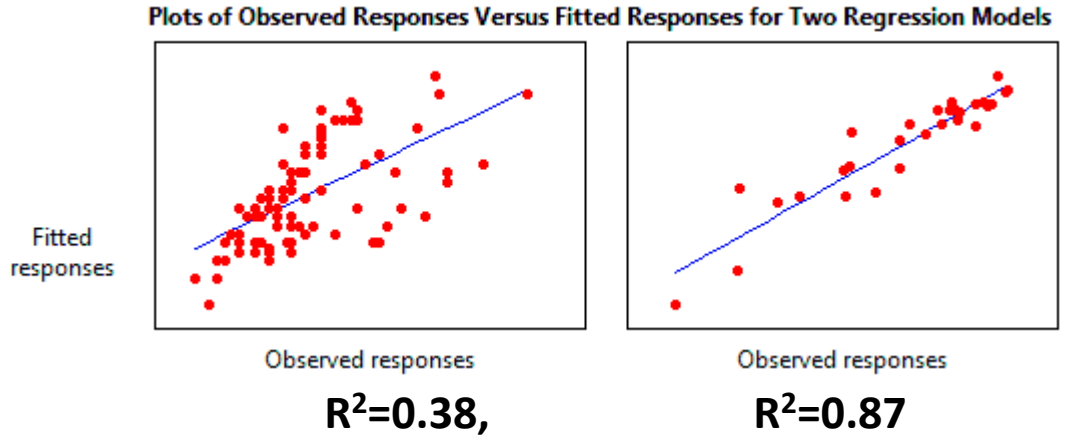
$$R^2=\frac{SSR}{SST}=\frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2}=1-\frac{\sum_i(y_i-\hat{y}_i)^2}{\sum_i(y_i-\bar{y})^2}$$

# 3.2.3 $R^2$ vs. R

$R^2$ 是两个变量的相关系数的平方。由于其含义和解释和R不同，常被单独用来作衡量两个变量之间相关性.  R-squared = Explained variation / Total variation

在多元线性回归中，R平方有更直接的解释，衡量**因变量的变异**中可**由自变量解释**的部分所占的比例.

例：在研究中$R^2$的使用的描述




**$R^2$=0.38,**  **$R^2$=0.87**

**Model Summary[b]**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .850[a] | .723 | .690 | 4.57996 |

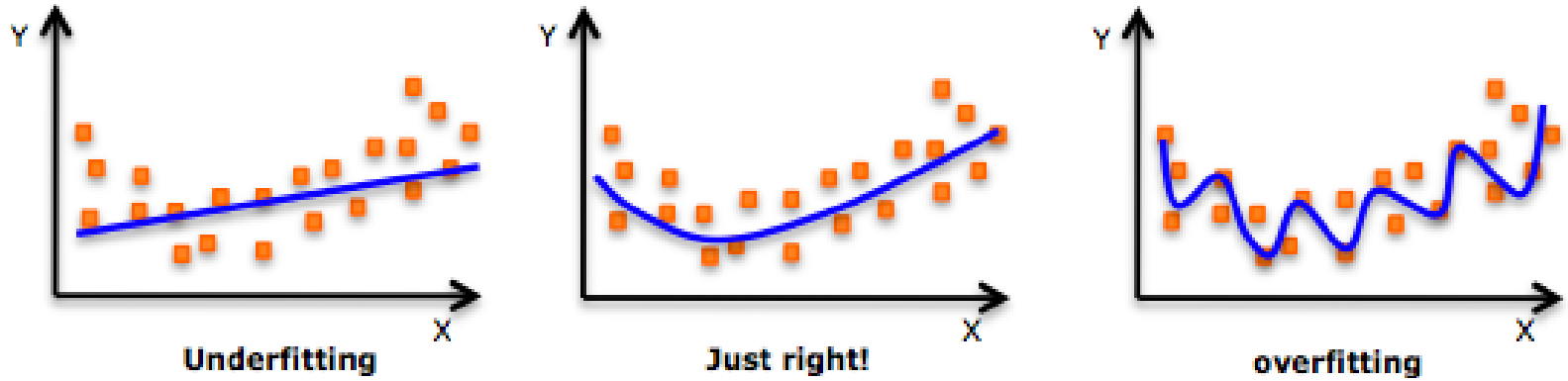
a. Predictors: (Constant), weight, horsepower

b. Dependent Variable: mpg

- Coefficient of Correlation is the R value *i.e.* 0.850
- Coefficient of Determination is the R square ($R^2$) value *i.e.* 0.723 (or 72.3%)
- R square is simply square of R i.e. R times R.

https://www.datasciencecentral.com/profiles/blogs/regression-analysis-how-do-i-interpret-r-squared-and-assess-the

# 3.2.4 用$R^2$来衡量 Underfitting vs overfitting of regression




**欠拟合（underfitting）：**

当模型无法准确捕获数据之间的依赖关系时（通常是由于模型自身的简单性），就会发生欠拟合。当应用于新数据时，它通常会**产生较低的 $R^2$，且泛化能力很差**

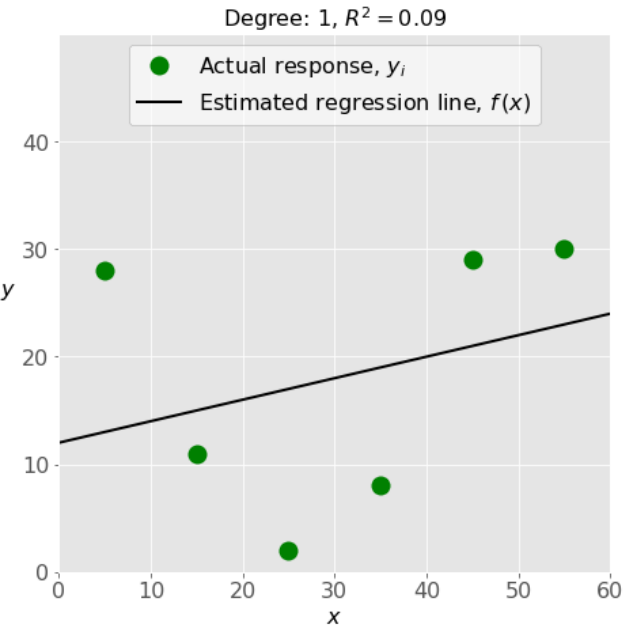
**过拟合（overfitting）：**

当模型同时学习数据之间的依赖关系和随机波动时，就会发生过拟合。换句话说，模型对现有数据学习能力太强。很多具有许多特别功能或术语的复杂模型通常易于过拟合。

当应**用于已知数据时**，此类模型通常会**产生较高的 $R^2$。但泛化能力不好，在用于新数据时，$R^2$ 就会显著降低**

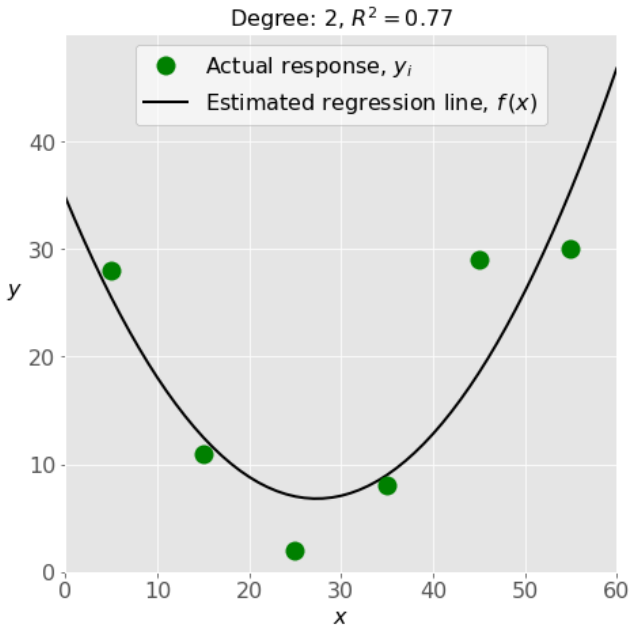
# 3.2.4 用$R^2$ 来衡量 Underfitting vs overfitting of regression

**$R^2$ = 0.09**



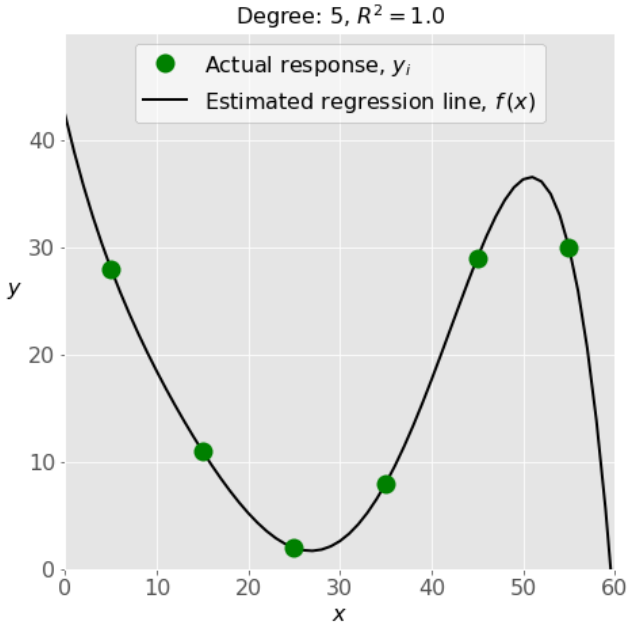

underfitting

**$R^2$ = 0.77**



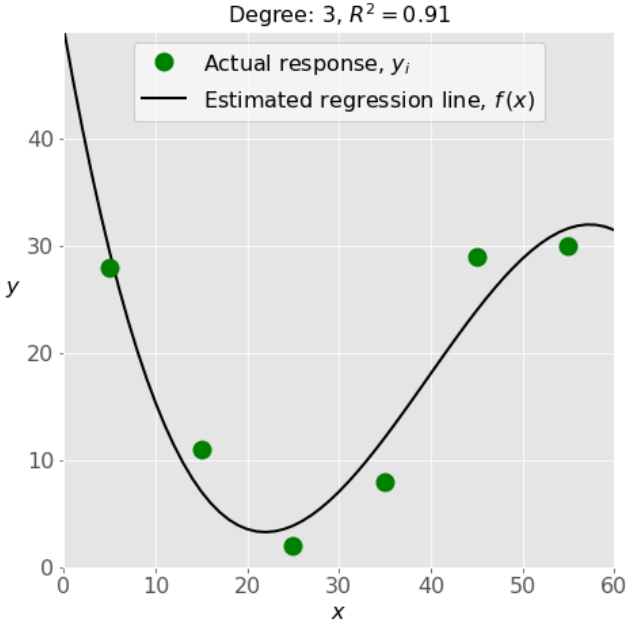

well-fitting

**$R^2$ = 1.0**




overfittings

**$R^2$ = 0.9**




overfittings

# 3.2.5 协方差，相关系数，回归模型 的 三者关系

两个独立变量 X, Y

有相关的两个变量 $y_i$ $\hat{y}_i$

协方差 → 相关系数（*R*）(特殊的协方差) → 回归模型（regression）拟合度 $R^2$

*Cov (X, Y)* → $\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = R$

Pearson R
Spearman R
Kendall τ

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

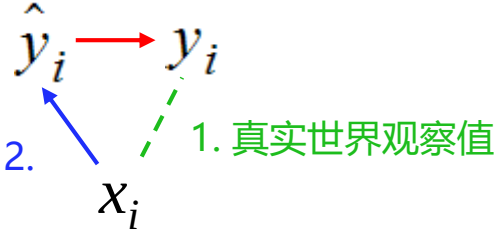
$y_i$ **：观察值（real value）**

$\bar{y}$ ：观察值平均值（mean of real value）

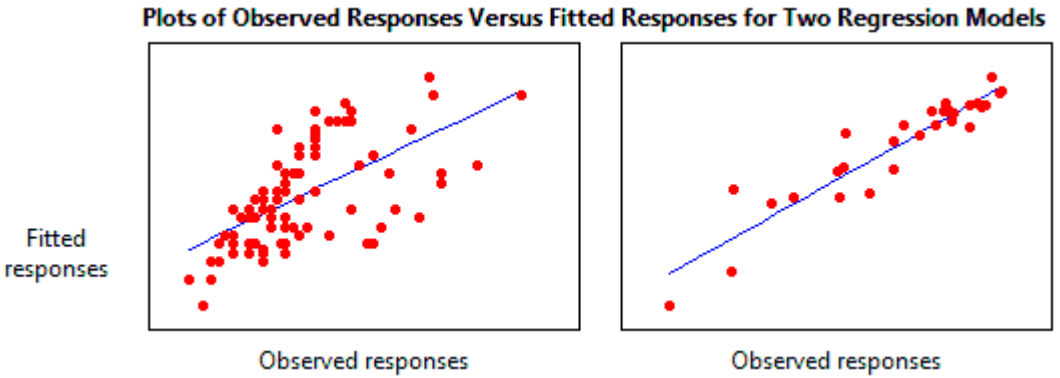
$\hat{y}_i$ **：模型预测值（predicted value from model）**

协同性 $x_i \longleftrightarrow y_i$ → 相关性 $x_i \longleftrightarrow y_i$ →

3. 拟合度



# 研究中$R^2$的使用描述例子




**$R^2$=0.38,** **$R^2$=0.87**

**Model Summary[b]**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .850[a] | .723 | .690 | 4.57996 |

a. Predictors: (Constant), weight, horsepower

b. Dependent Variable: mpg

- Coefficient of Correlation is the **R value *i.e.* 0.850**
- Coefficient of Determination is the **R square ($R^2$)** value *i.e.* **0.723** (or 72.3%)
- R square is simply square of R i.e. R times R.

https://www.datasciencecentral.com/profiles/blogs/regression-analysis-how-do-i-interpret-r-squared-and-assess-the

# 3.2.6 Adjusted R-squared (调整R方)

- R-squared（值范围0-1）描述的输入变量对输出变量的解释程度。在单变量（一元）线性回归中，R-squared 越大，说明拟合程度越好。
- **R-squared 存在的问题：**
  - 问题1：你给一个模型每加一个新变量/预测因子，R方就会增加，即使新变量可能和模型没啥关系。 R方从不减少。因此，**有时一个多变量的模型适合度看起来不错，可能仅仅因为比别人有更多变量而已。**
  - 问题2：如果一个模型有太多的预测项和高阶多项式，它就开始会对数据中的随机噪声进行建模。这种情况被称为过度拟合模型，它会产生误导性的高R平方值，并会降低预测能力。

# 3.2.6 Adjusted R-squared (调整R方)

- R-squared（值范围0-1）描述的输入变量对输出变量的解释程度。在单变量（一元）线性回归中，R-squared 越大，说明拟合程度越好。
- **R-squared 存在的问题：**
  - 问题1：你给一个模型每加一个新变量/预测因子，R方就会增加，即使新变量可能和模型没啥关系。 R方从不减少。因此，**有时一个多变量的模型适合度看起来不错，可能仅仅因为比别人有更多变量而已。**
  - 问题2：如果一个模型有太多的预测项和高阶多项式，它就开始会对数据中的随机噪声进行建模。这种情况被称为过度拟合模型，它会产生误导性的高R平方值，并会降低预测能力。
- 需要 Adjusted R-squared：对增加一个不会改善模型效果的变量的模型进行惩罚（不是越多变量越好）
- 结论：如果**单变量**线性回归，则使用 R-squared 评估，**多变量**则使用 adjusted R-squared
  - 单变量线性回归中，R-squared 和 adjusted R-squared 是一致的
  - 如果增加更多无意义的变量，则 R-squared 和 adjusted R-squared 间的差距会越来越大，adjusted R-squared会下降。但是如果加入的特征值是显著的，则adjusted R-squared也会上升

The formula is:

$$R^2_{adj} = 1 - \left[ \frac{(1 - R^2)(n-1)}{n-k-1} \right]$$

where:

- N is the number of points in your data sample.
- K is the number of independent regressors, i.e. the number of variables in your model, excluding the constant.

| Vars | R-Sq | R-Sq(adj) |
|---|---|---|
| 1 | 72.1 | 71.0 |
| 2 | 85.9 | 84.8 |
| 3 | 87.4 | 85.9 |
| 4 | 89.1 | 82.3 |
| 5 | 89.9 | 80.7 |

# 3.2.7 Predicted R-squared (预测R方/预测拟合度)

- 使用预测$R^2$来确定回归模型的**预测效果**。
  - 这一统计量有助于识别一个模型是否有存在：**对训练数据能比较好的拟合，但对新数据却不能有效预测的情况。**即使你不使用你的模型来对新数据进行预测，预测拟合度仍然可以提供关于你的模型的有价值的见解。

Statistical software calculates predicted R-squared using the following procedure:

- It removes a data point from the dataset.
- Calculates the regression equation.
- Evaluates how well the model predicts the missing observation.
- And, repeats this for all data points in the dataset.

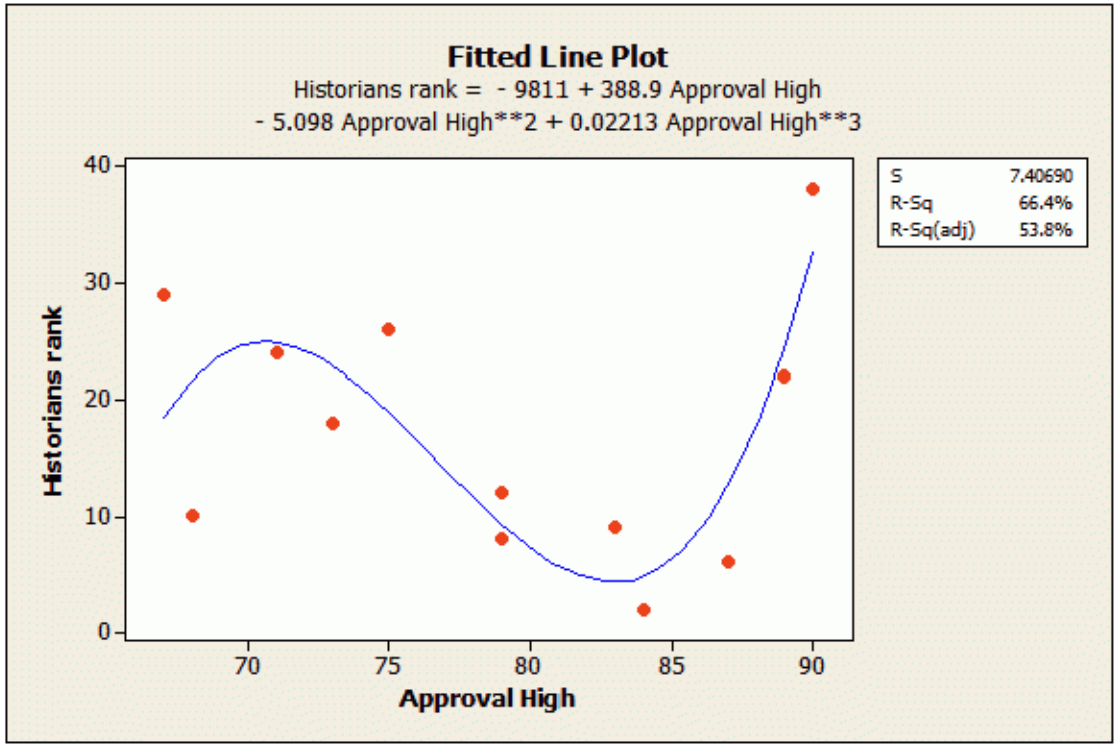
# 3.2.7 Predicted R-squared (预测拟合度)

- 使用预测$R^2$来确定回归模型的**预测效果**。
  - 这一统计量有助于识别一个模型是否有存在：**对训练数据能比较好的拟合，但对新数据却不能有效预测的情况。**即使你不使用你的模型来对新数据进行预测，预测拟合度仍然可以提供关于你的模型的有价值的见解。

Statistical software calculates predicted R-squared using the following procedure:

- It removes a data point from the dataset.
- Calculates the regression equation.
- Evaluates how well the model predicts the missing observation.
- And, repeats this for all data points in the dataset.

- 预测$R^2$ 有助于确定回归模型是否过度拟合。
  - 过拟合模型会包含了过多的变量和参数，因为它开始拟合样本中的随机噪声。
- 根据其定义，模型不应该去预测随机噪声。因此，如果一个模型能拟合大量的随机噪声，则预测$R^2$值必会下降。如果预测的$R^2$明显小于$R^2$，这就是个警告信号，表明你已经过拟合了模型。应尝试减少模型里使用的预测因子/变量/参数的个数。



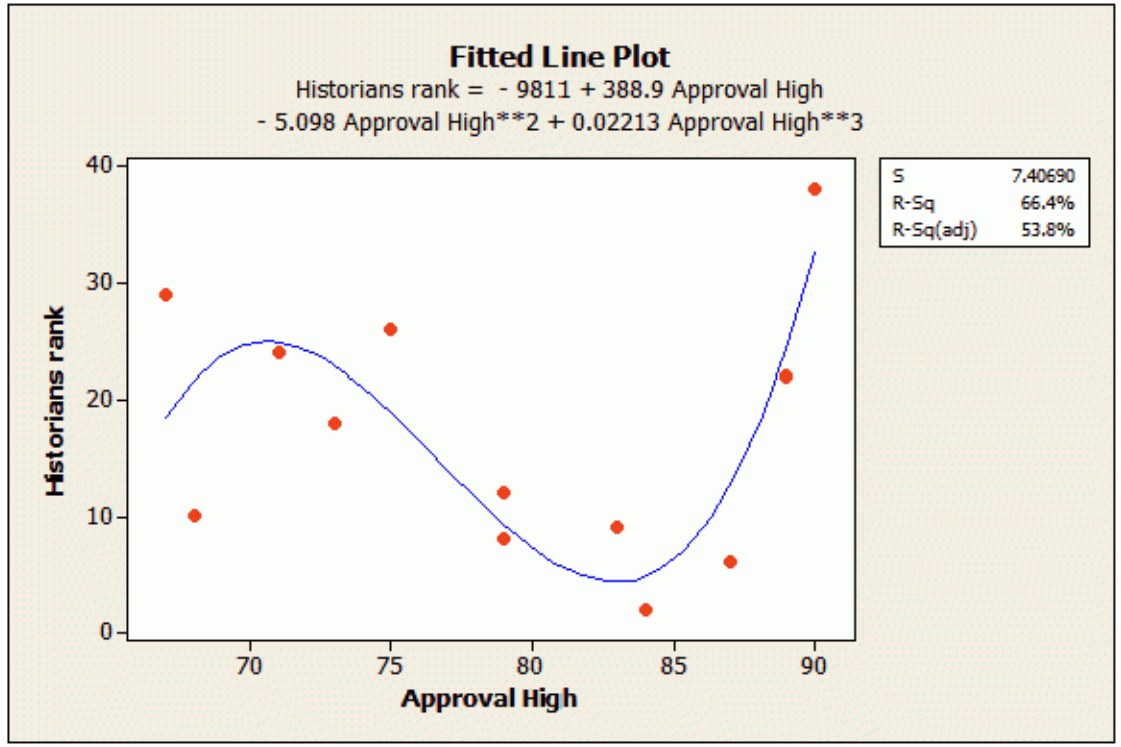

## Regression Analysis: Historians rank versus Approval High

Analysis of Variance

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Regression | 3 | 867.10 | 289.034 | 5.27 | 0.027 |
| Approval High | 1 | 438.35 | 438.347 | 7.99 | 0.022 |
| Approval High*Approval High | 1 | 460.23 | 460.225 | 8.39 | 0.020 |
| Approval High*Approval High*Approval High | 1 | 481.55 | 481.552 | 8.78 | 0.018 |
| Error | 8 | 438.90 | 54.862 | | |
| Lack-of-Fit | 7 | 430.90 | 61.557 | 7.69 | 0.271 |
| Pure Error | 1 | 8.00 | 8.000 | | |
| Total | 11 | 1306.00 | | | |

Model Summary

| S | R-sq | R-sq(adj) | R-sq(pred) |
|---|---|---|---|
| 7.40690 | 66.39% | 53.79% | 0.00% |




**Regression Analysis: Historians rank versus Approval High**

Analysis of Variance

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Regression | 3 | 867.10 | 289.034 | 5.27 | 0.027 |
| Approval High | 1 | 438.35 | 438.347 | 7.99 | 0.022 |
| Approval High*Approval High | 1 | 460.23 | 460.225 | 8.39 | 0.020 |
| Approval High*Approval High*Approval High | 1 | 481.55 | 481.552 | 8.78 | 0.018 |
| Error | 8 | 438.90 | 54.862 | | |
| Lack-of-Fit | 7 | 430.90 | 61.557 | 7.69 | 0.271 |
| Pure Error | 1 | 8.00 | 8.000 | | |
| Total | 11 | 1306.00 | | | |

Model Summary

| S | R-sq | R-sq(adj) | R-sq(pred) |
|---|---|---|---|
| 7.40690 | 66.39% | 53.79% | 0.00% |

https://blog.minitab.com/blog/adventures-in-statistics-2/multiple-regession-analysis-use-adjusted-r-squared-and-predicted-r-squared-to-include-the-correct-number-of-variables

**点评：**

**R平方**和**调整后的R平方**，看起来都很好，因为它们的p值都小于0.05。

**但**其实我们只是扭曲了回归线，迫使它把每个点都连起来，而不是找到一个实际的关系。

我们对模型过度拟合了，预测的R平方为0%直接就说明了这个模型的预测结果不行。


Adjusted R-squared and predicted R-square help you resist the urge to add too many independent variables to your model.

- Adjusted R-square compares models with different numbers of variables.
- Predicted R-square can guard against models that are too complicated.

# 谢谢，下周见！