《生物实验设计》
第四章　统计推断

王超

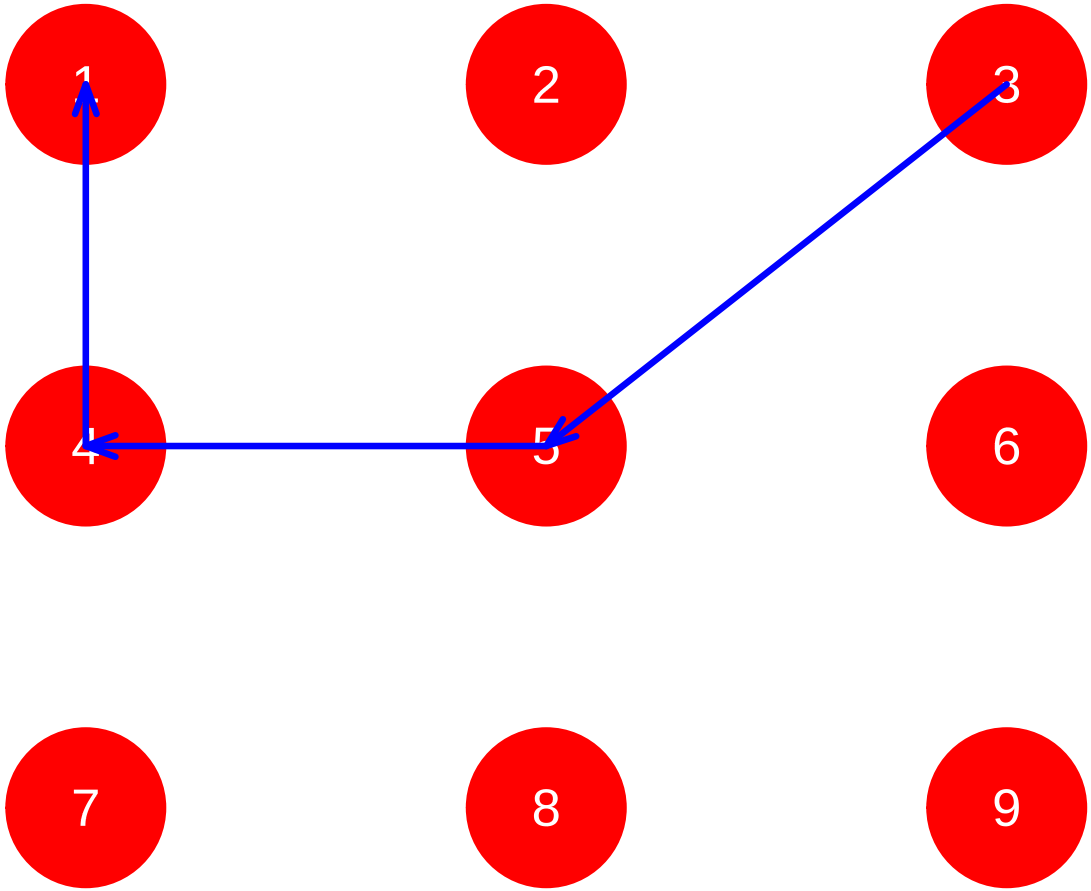
广东药科大学

Email: wangchao@gdpu.edu.cn

2022-09-26

第七章　直线回归与相关分析

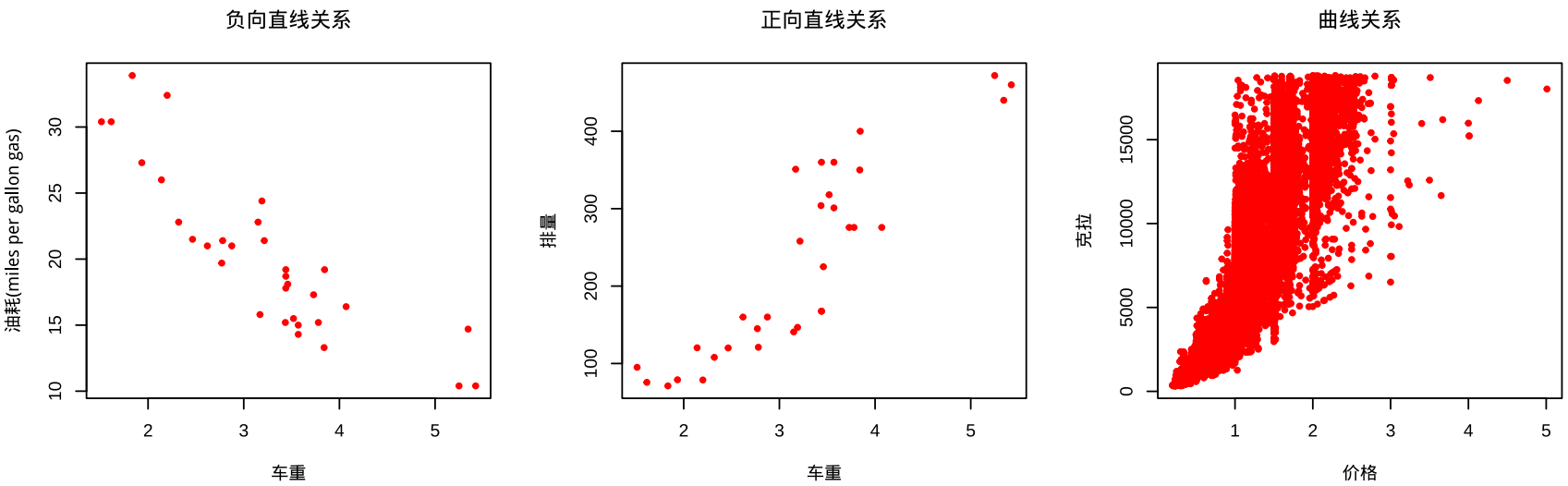
Check In



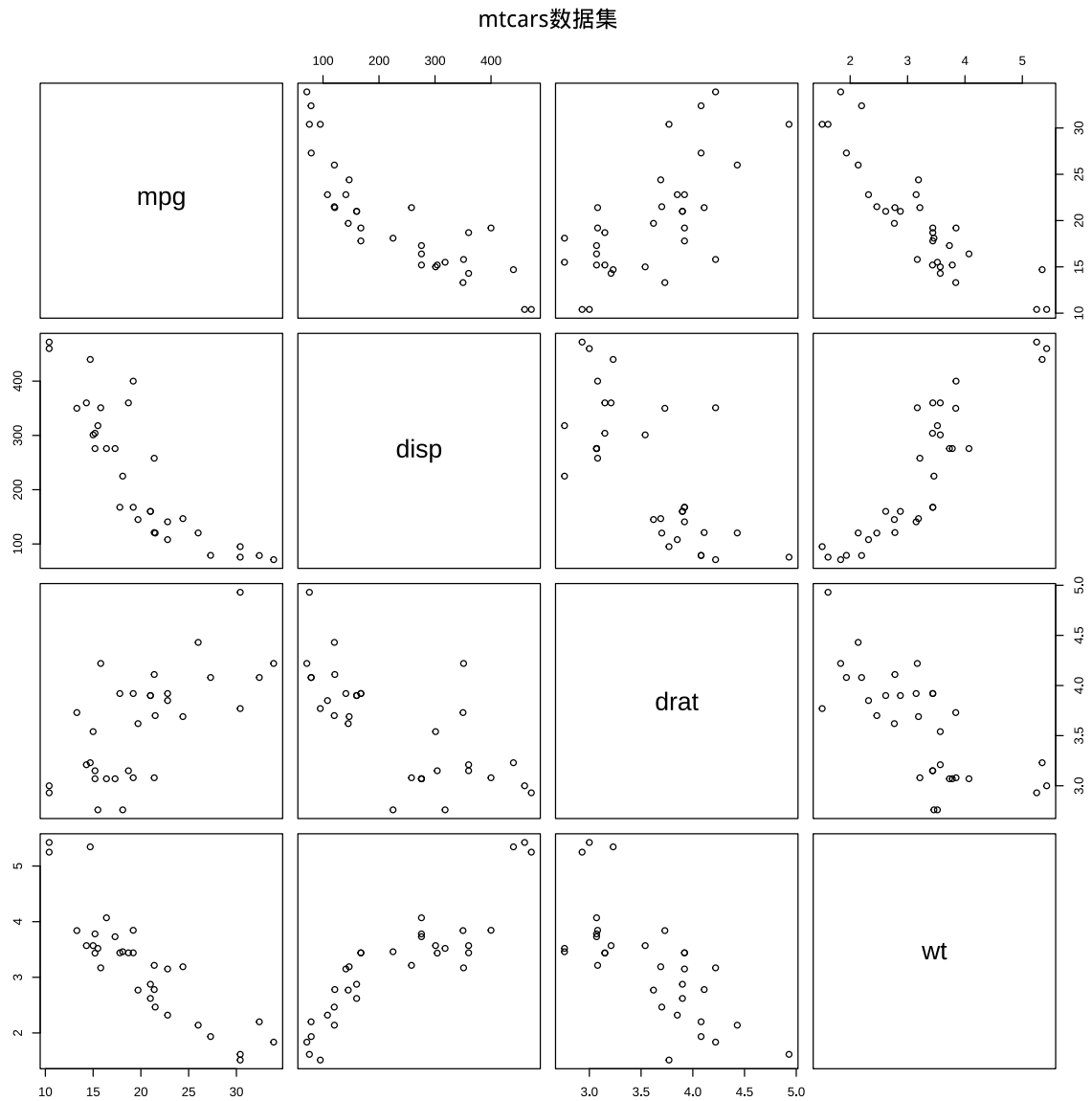
第一节　回归和相关的概念

- 变量间的相互关系：
  - 因果关系
    - 一个变量的变化受另一个变量或几个变量的制约
  - 平行关系
    - 两个以上变量之间共同受到另外因素的影响
- 两个变量的成对观测值可表示为
  $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$
- 每对观测值在平面直角坐标系中表示成一个点，作成散点图

- 散点图



- 从散点图可以看到：
  - 变量间关系的性质和程度
  - 变量间关系的类型
  - 是否有异常值干扰


mtcars数据集
mpg
disp
drat
wt

- 因果关系
  - 用回归分析研究
  - 自变量 x，因变量 y
  - 因变量随着自变量的变化而变化，具有随机误差
  - 回归关系
- 一元回归分析
  - 一个自变量与一个因变量
  - 直线回归
  - 曲线回归
- 多元回归分析
  - 多个自变量与一个因变量
- 揭示因果关系的变量之间的联系形式，建立回归方程，利用回归方程预测和控制因变量

- 平行关系
  - 用相关分析研究
  - 变量 x 和变量 y 无自变量和因变量之分，都具有随机误差
  - 相关关系
- 直线相关分析
  - 两个变量的直线关系
- 复相关分析
  - 一个变量与多个变量间的线性相关
- 偏相关分析
  - 其余变量保持不变的情况下两个变量间的线性相关
- 研究两个变量之间相关的程度和性质或一个变量与多个变量之间相关的程度

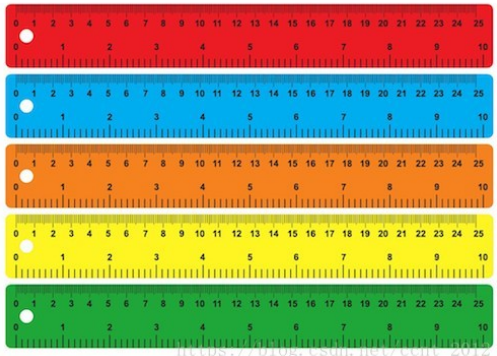
- 对于自变量 x 的每一个取值 x_i，都有因变量 y 的一个分布与之对应
- 条件平均数
  - 当 $x = x_i$ 时，y_i 的平均数 μ_{y_i} 与之对应
- 利用直线回归方程描述这种关系：
  - $\hat{y} = a + bx$
  - a 为截距，b 为系数，$\hat{y}$ 为因变量 y 的点估计

- 两个变量呈线性关系，可以用直线回归来描述
- 最小二乘法
  - 解决曲线拟合问题最常用的方法
  - 基本思路是求 a, b, 令因变量的观测值与回归估计值的离均差平方和 Q 值最小

$$min(Q) = \sum_{1}^{n}(y - \hat{y})^2 = \sum_{1}^{n}(y - a - bx)^2$$

  - 天体运动论，1809，高斯
  - 计算谷神星轨道
  - 通过最小化误差的平方和寻找数据的最佳函数匹配

用五把不同颜色的尺子分别测量一线段的长度，得到的数值分别为：

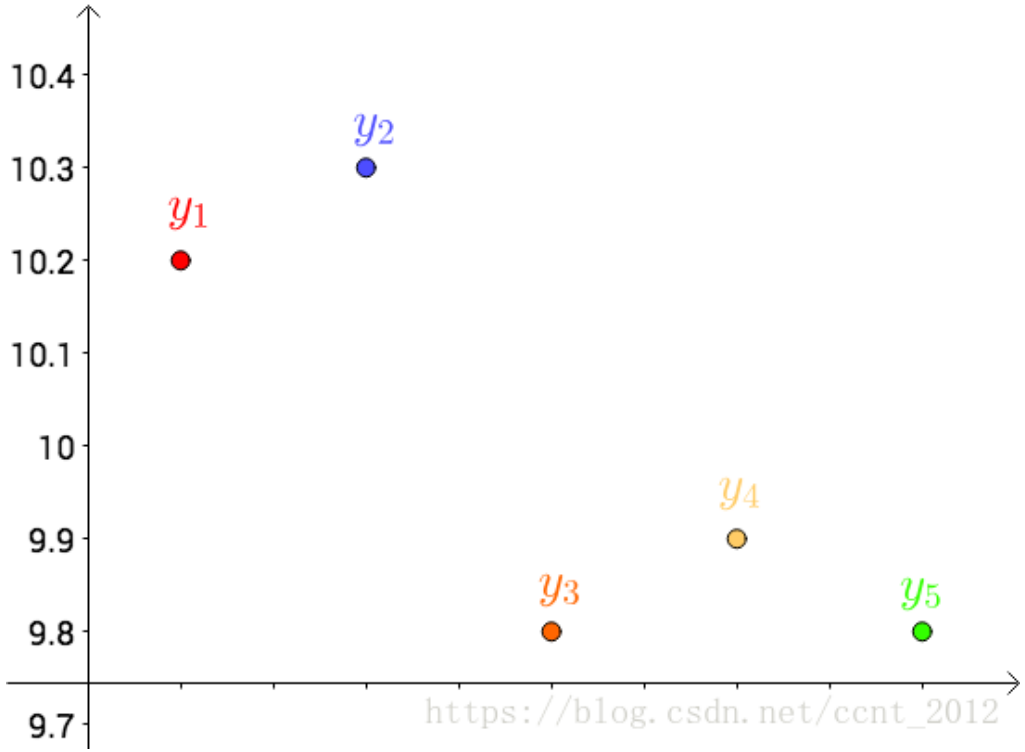


红	蓝	橙	黄	绿
10.2	10.3	9.8	9.9	9.8

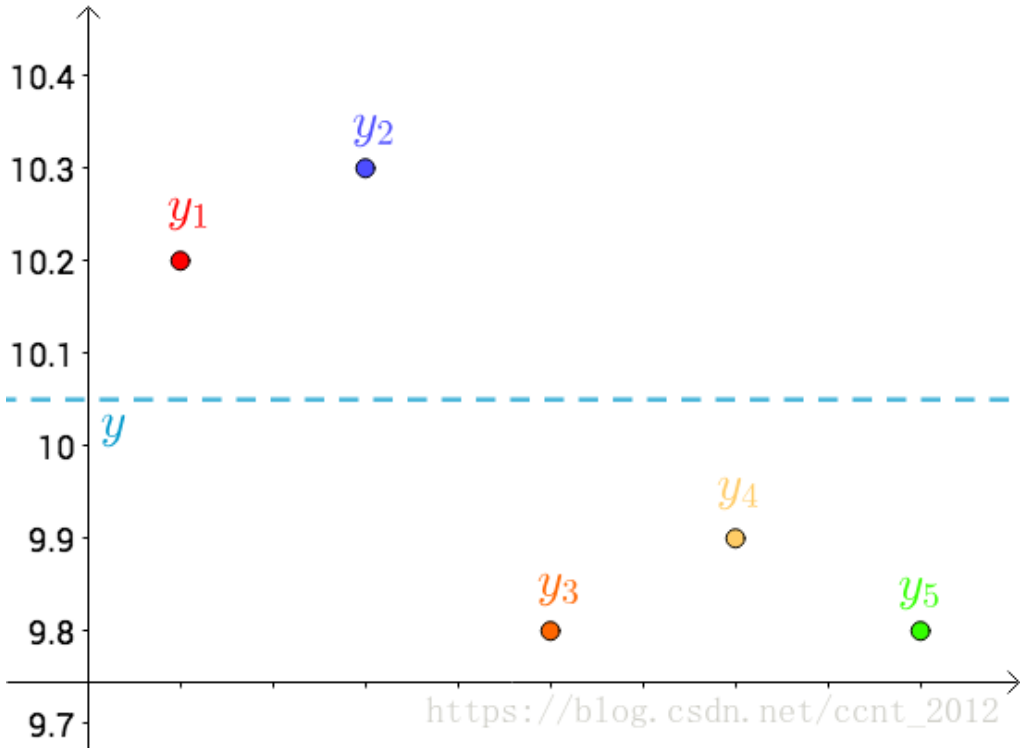
一般用平均值来作为线段长度：

$$\bar{x} = \frac{10.2 + 10.3 + 9.8 + 9.9 + 9.8}{5} = 10$$

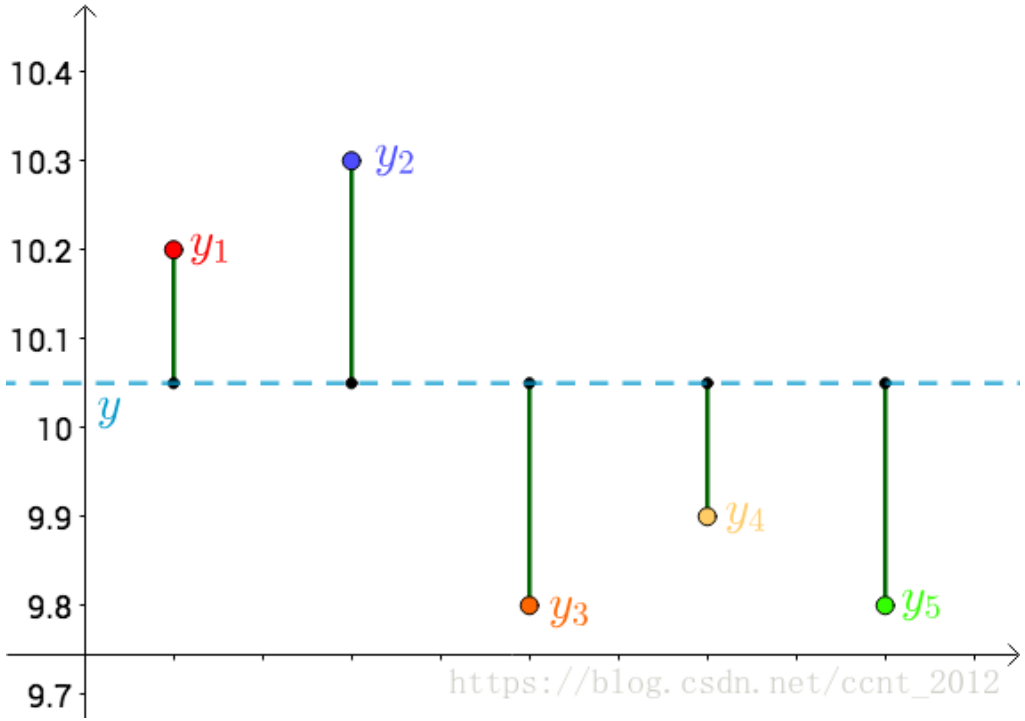
把测试得到的值画在坐标系中，分别记作 y_i



把要猜测的线段长度的真实值用平行于横轴的直线来表示，分别记作 y



每个点都向 y 做垂线，垂线的长度就是 $y - y_i$，也可以理解为测量值和真实值之间的误差：



- 因为误差是长度，还要取绝对值，计算起来麻烦，就干脆用平方来代表误差：

$$|y - y_i| \Rightarrow (y - y_i)^2$$

- 总的误差平方就是

$$\sigma = \sum (y - y_i)^2$$

- 因为 y 是猜测的，所以可以上下不断变换
- 方差不断变化

- 勒让德（Adrien-Marie Legendre）提出让总的误差的平方最小的 y 就是真值，这是基于:
  - 如果误差是随机的，应该围绕真值上下波动
  - 无偏估计

$$\sigma = min \sum (y - y_i)^2$$

- 对二次函数求导:

$$\frac{d}{dy}\sigma = \frac{d}{dy}\sum (y-y_i)^2 = 2\sum (y-y_i) = 2((y-y_1)+(y-y_2)+\ldots+(y-y_5)) = 0$$

$$5y = y_1 + y_2 + \ldots + y_5 \Rightarrow y = \frac{y_1 + y_2 + \ldots + y_5}{5}$$

- 求真值的最小二乘法

$$\sigma = min \sum (y - y_i)^2$$

- 求自变量和因变量关系的最小二乘法

$$min(Q) = \sum_1^n (y - \hat{y})^2 = \sum_1^n (y - a - bx)^2$$

  - 根据极值定理，对 a 和 b 分别求导：

$$\frac{\partial Q}{\partial a} = -2 \sum (y - a - bx) = 0, \frac{\partial Q}{\partial b} = -2 \sum (y - a - bx)x = 0$$

  - 整理得到：

$$\begin{cases} an + b \sum x = \sum y \\ a \sum x + b \sum x^2 = \sum xy \end{cases}$$

- 最后得到：

$$\begin{cases} a = \bar{y} - b\bar{x} \\ b = \dfrac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} \end{cases}$$

- $a > 0$，回归直线在第一象限与 y 轴相交
- $a < 0$，回归直线在第一象限与 x 轴相交
- $b > 0$，y 随 x 的增加而增加
- $a < 0$，y 随 x 的增加而减小