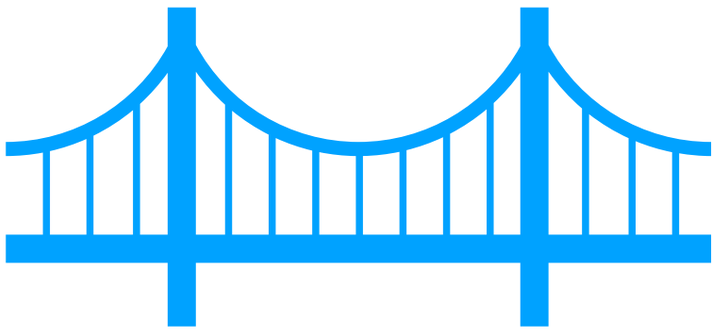
定量分析：数据思维与商业统计

陈文波
cwb@whu.edu.cn
2021年10月

课程梗概

描述性分析与EDA

- 数值方法：三类指标
- 图形方法：五种图形



- 抽样分布、CLT
- 正态分布
- T分布
- 卡方分布
- F分布

推断性分析

- 区间估计
- 假设检验
- 应用：
  - 分类 vs. 分类
  - 分类 vs. 数值
  - 数值 vs. 数值

R

课程梗概

- Session 1：内容简介、基本概念、描述性统计1
- Session 2：描述性统计2
- Session 3：抽样、中心极限定理、正态分布
- Session 4：其他三个分布，区间估计
- Session 5：假设检验，两个总体均值和比例的推断
- Session 6：总体方差和方差比的推断、分类数据 Vs. 分类数据： 卡方检验与列联表分析
- Session 7：分类数据 Vs. 数值型数据： 方差分析；数值型数据：回归分析
- Session 8：课程总结

统计推断的应用

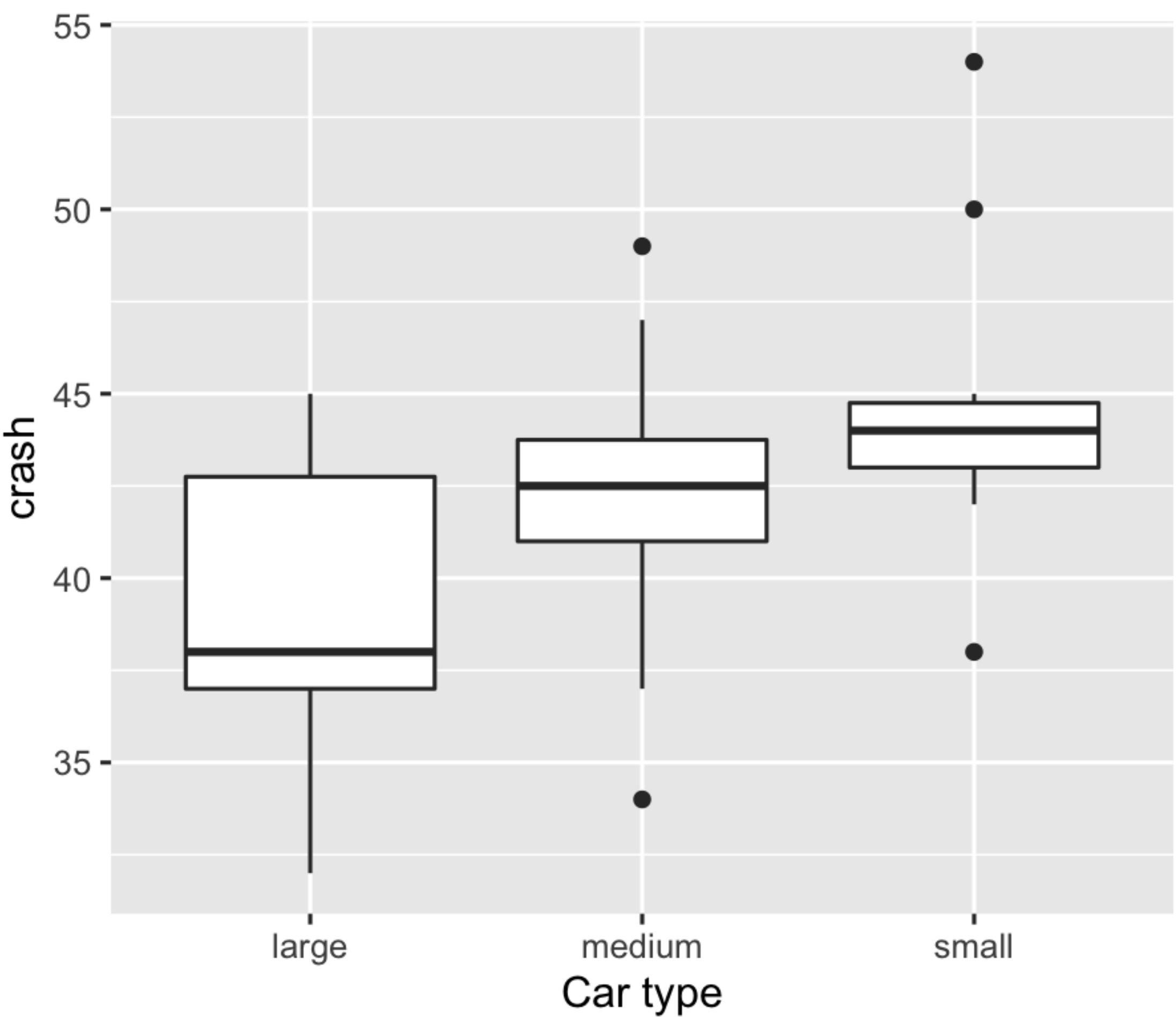
x		y
分类	χ^2	分类
分类	F ANOVA	数值
数值	logit	分类
数值	lm	数值

方差分析：实际问题

- 大🚕更安全吗?

Table 12-1 Chest Deceleration Measurements (in g) from Car Crash Tests

Small Cars	44	43	44	54	38	43	42	45	44	50	$\rightarrow \bar{x} = 44.7$ g
Medium Cars	41	49	43	41	47	42	37	43	44	34	$\rightarrow \bar{x} = 42.1$ g
Large Cars	32	37	38	45	37	33	38	45	43	42	$\rightarrow \bar{x} = 39.0$ g


55
50
45
40
35
crash
large
medium
small
Car type

检验

$$F = \frac{\text{variance between samples}}{\text{variance within samples}} = \frac{\left[\dfrac{\Sigma n_i(\bar{x}_i - \bar{\bar{x}})^2}{k-1}\right]}{\left[\dfrac{\Sigma(n_i - 1)s_i^2}{\Sigma(n_i - 1)}\right]}$$

where

$\bar{\bar{x}}$ = mean of all sample values combined

k = number of population means being compared

n_i = number of values in the ith sample

$\bar{x}_i$ = mean of values in the ith sample

s_i^2 = variance of values in the ith sample

到底是哪两个不相等?

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\text{MS(error)} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

双因素方差分析

Table 12-3 Chest Deceleration Measurements (in g) from Car Crash Tests

	Size of Car		
	Small	Medium	Large
Foreign	44	41	32
	54	49	45
	43	47	42
Domestic	43	43	37
	44	37	38
	42	34	33

Wrap-up

- 在二维空间解决一维空间的问题
- 方差分解思想
- 单因素与双因素方差分析

练习：红酒配红肉

		红色肉类	鸡肉	鱼
红葡萄酒	1	10	4	3
	2	9	4	2
	3	10	2	4
白葡萄酒	4	3	6	8
	5	3	5	6
	6	2	7	7

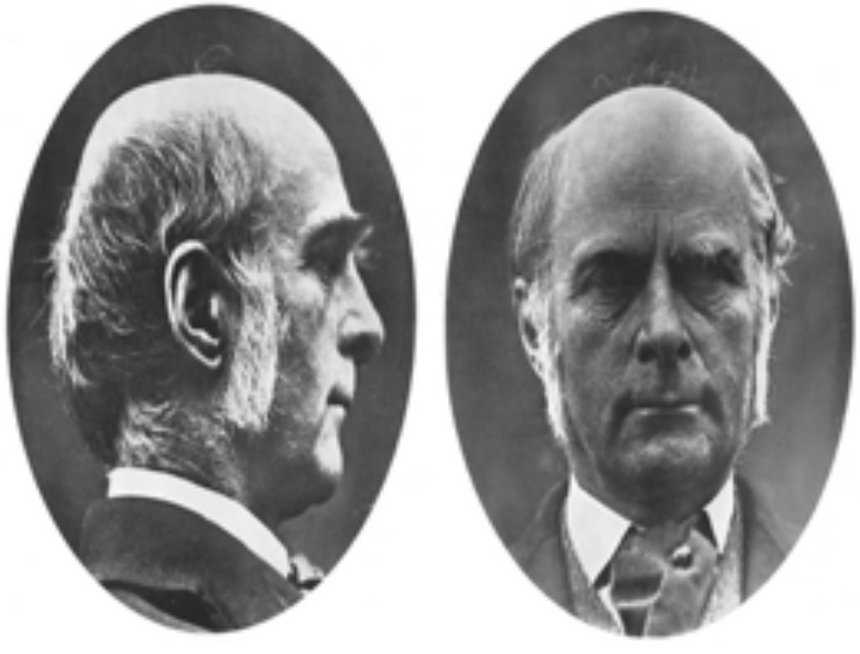
相关与回归分析

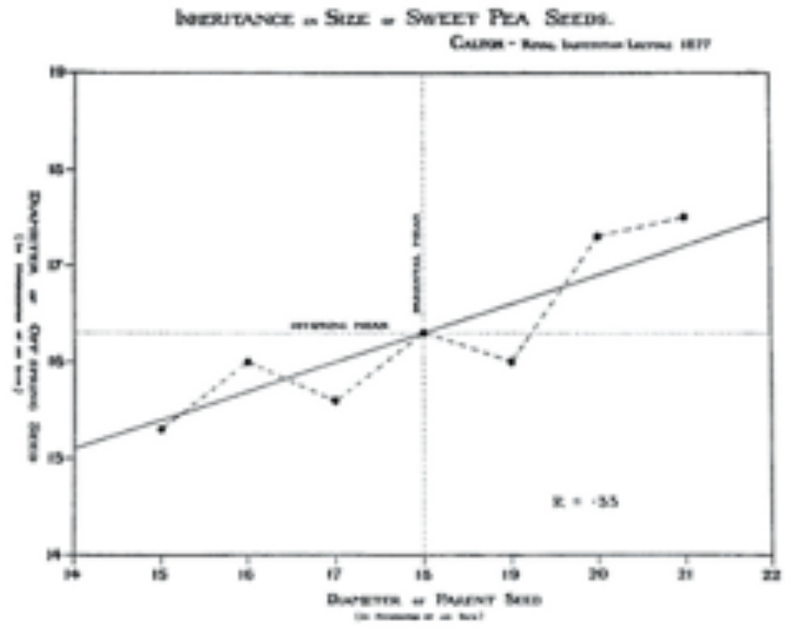
- 相关与线性相关系数

回归分析

Sir Francis Galton (1822-1911)

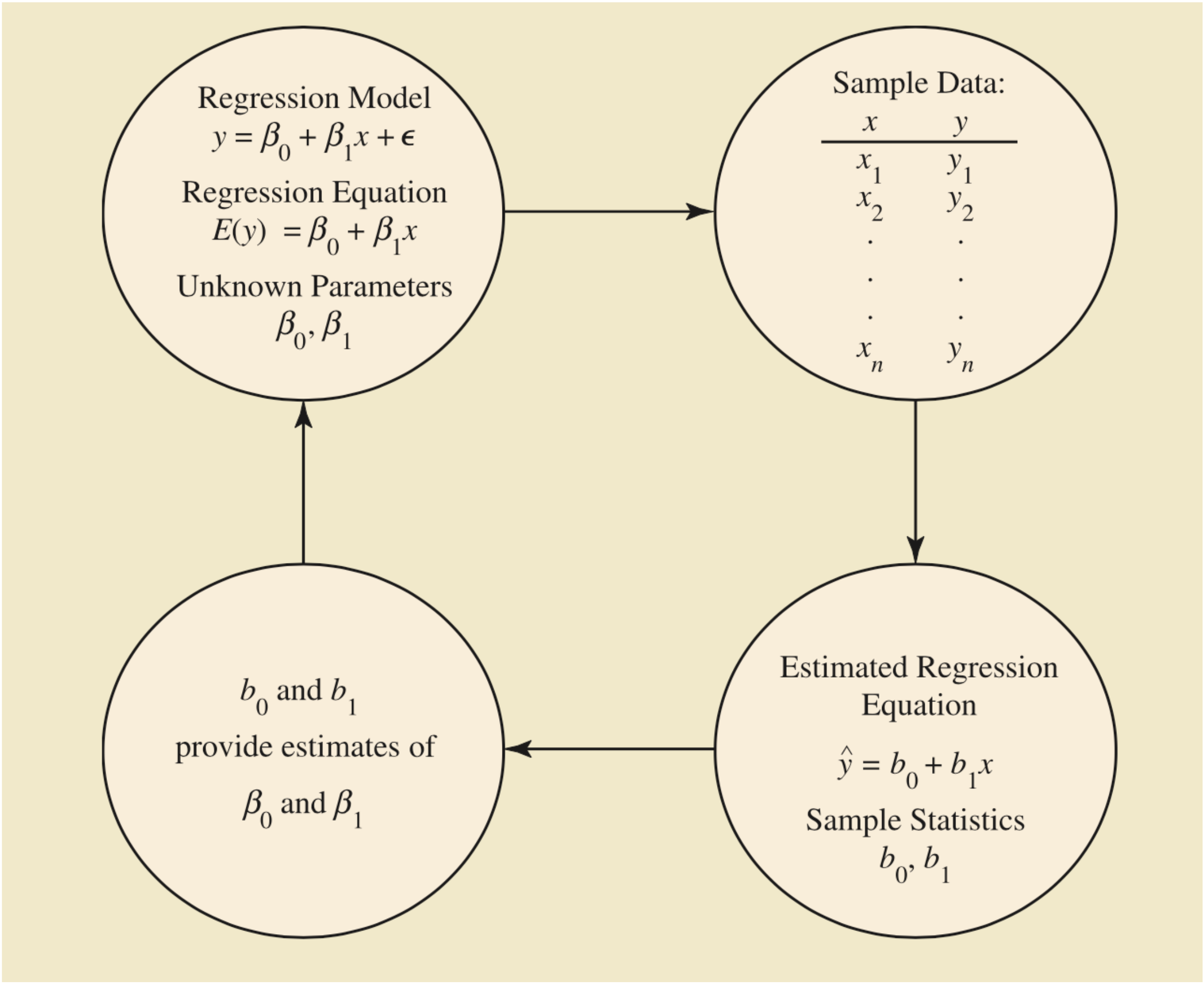
- Introduced correlation and regression analysis for hereditary research
- Worlds first regression line!
- Coined the term "Regression to the mean"



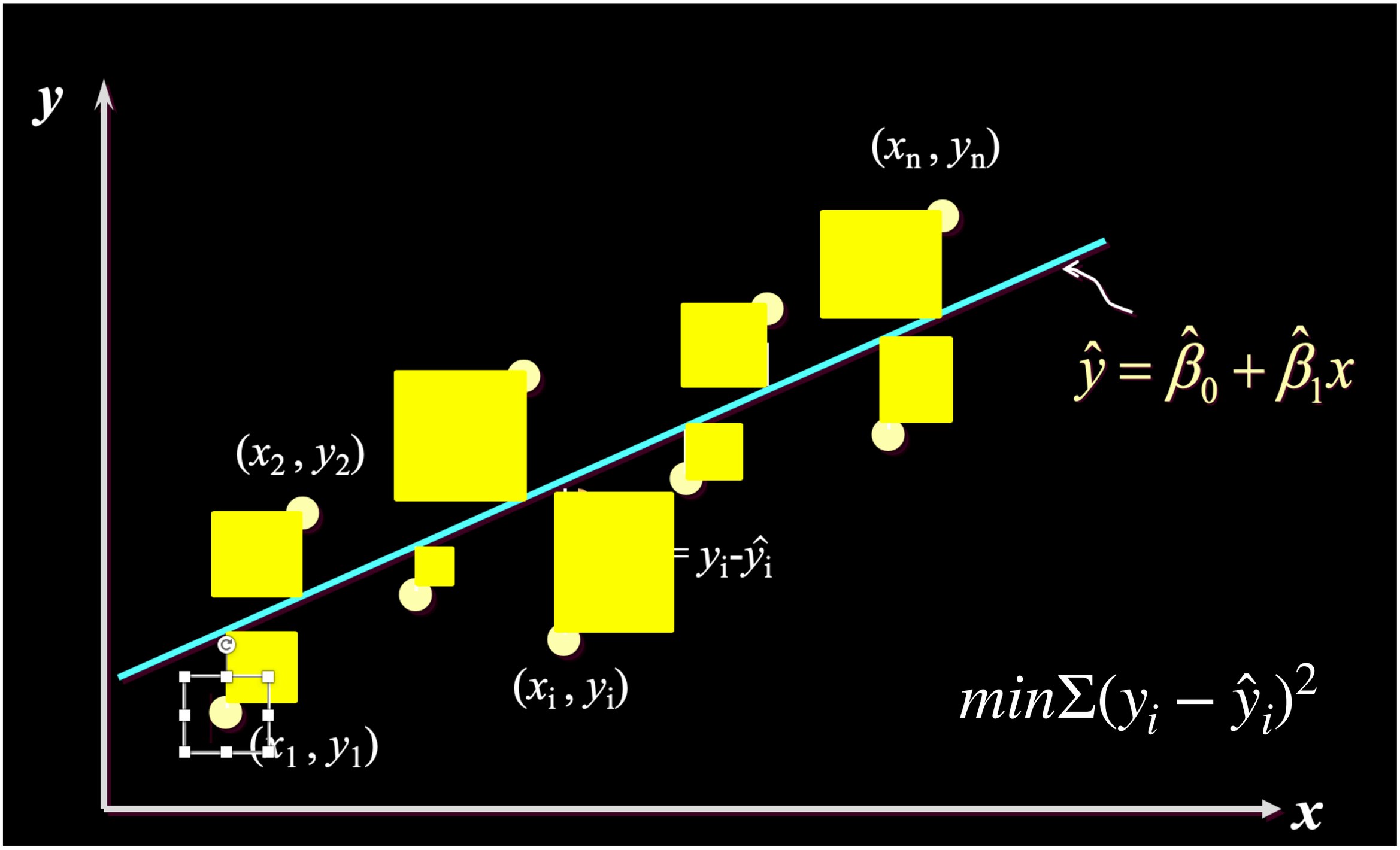


回归模型、回归方程

- 回归模型：$y = \beta_0 + \beta_1 x + \epsilon$
- 回归方程：$E(y) = \beta_0 + \beta_1 x$
- 估计的回归方程：$\hat{y} = b_0 + b_1 x$


Regression Model
$y = \beta_0 + \beta_1 x + \epsilon$
Regression Equation
$E(y) = \beta_0 + \beta_1 x$
Unknown Parameters
β_0, β_1
Sample Data:
x y
x_1 y_1
x_2 y_2
x_n y_n
Estimated Regression Equation
$\hat{y} = b_0 + b_1 x$
Sample Statistics
b_0, b_1
b_0 and b_1
provide estimates of
β_0 and β_1

最小二乗法


y
(x_n, y_n)
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
(x_2, y_2)
$= y_i - \hat{y}_i$
(x_i, y_i)
(x_1, y_1)
$min\Sigma(y_i - \hat{y}_i)^2$
x

系数

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

where

x_i = value of the independent variable for the ith observation

y_i = value of the dependent variable for the ith observation

$\bar{x}$ = mean value for the independent variable

$\bar{y}$ = mean value for the dependent variable

n = total number of observations

An alternate formula for b_1 is

$$b_1 = \frac{\Sigma x_i y_i - (\Sigma x_i \Sigma y_i)/n}{\Sigma x_i^2 - (\Sigma x_i)^2/n}$$

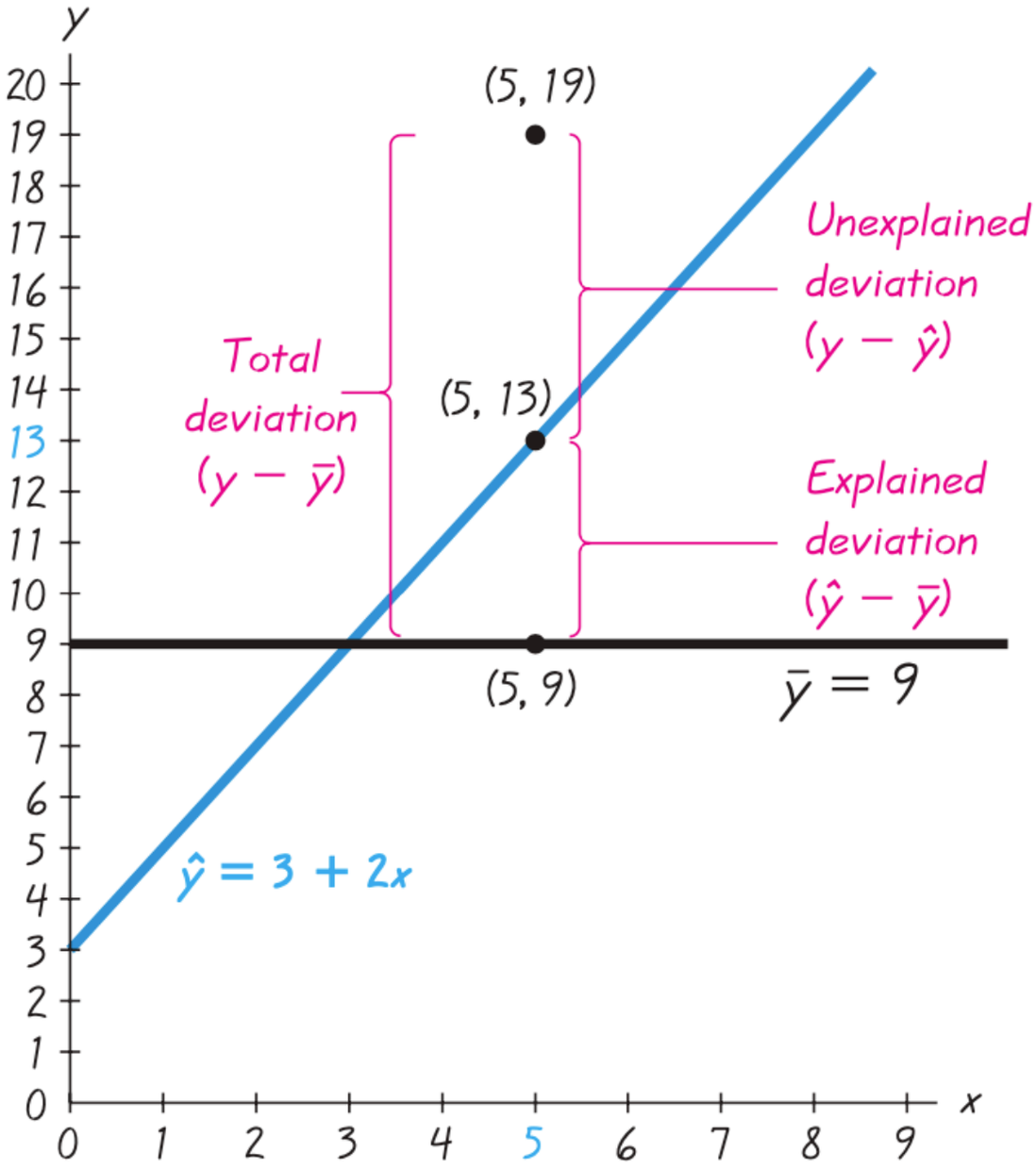
练习

- 对五个样本点的观测，其两个变量x,y的取值如下：

x_i	1	2	3	4	5
y_i	3	7	5	11	14

- 练习：
  - 画出x-y散点图
  - 从散点图可以看出x-y大致是什么关系?
  - 估计线性回归方程的两个系数
  - 用估计的回归方程预测当x=4时，y的取值。

判定系数：coefficient of determination



$$SSE = \Sigma(y_i - \hat{y}_i)^2$$

$$SST = \Sigma(y_i - \bar{y})^2$$

$$SSR = \Sigma(\hat{y}_i - \bar{y})^2$$

$$r^2 = \frac{SSR}{SST}$$

模型的前提假定：ϵ

1. ϵ是随机的，并且期望值是0

2. 对于所有的x, 其方差是一样的

3. ϵ的取值独立；

4. ϵ是正态分布的。

显著性检验

MEAN SQUARE ERROR (ESTIMATE OF σ^2)

$$s^2 = \text{MSE} = \frac{\text{SSE}}{n-2}$$

STANDARD ERROR OF THE ESTIMATE

$$s = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n-2}}$$

显著性检验： 回归系数

SAMPLING DISTRIBUTION OF b_1

Expected Value

$$E(b_1) = \beta_1$$

Standard Deviation

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\Sigma(x_i - \bar{x})^2}}$$

Distribution Form

Normal

ESTIMATED STANDARD DEVIATION OF b_1

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}}$$

显著性检验： 回归系数

t TEST FOR SIGNIFICANCE IN SIMPLE LINEAR REGRESSION

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

TEST STATISTIC

$$t = \frac{b_1}{s_{b_1}}$$

REJECTION RULE

p-value approach:	Reject H_0 if *p*-value $\leq \alpha$
Critical value approach:	Reject H_0 if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$

where $t_{\alpha/2}$ is based on a *t* distribution with $n - 2$ degrees of freedom.

显著性检验： 模型整体

TEST STATISTIC

$$F = \frac{\text{MSR}}{\text{MSE}} \qquad (14.21)$$

REJECTION RULE

p-value approach: Reject H_0 if p-value $\leq \alpha$

Critical value approach: Reject H_0 if $F \geq F_\alpha$

where F_α is based on an F distribution with 1 degree of freedom in the numerator and $n - 2$ degrees of freedom in the denominator.

其他问题

- 前提假设是否满足
- 特殊点：
  - 离群点
  - 高影响点
  - 杠杆点

回归分析步骤

1. 画散点图，看看：（1）是否大体是线性关系；（2）是否有离群点；

2. 进行线性拟合；

3. 画残差图（x-residual)，看看残差分布是不是没有模式（除了一条直线以外）；

4. 画残差的直方图或QQ图，看残差是否大致正态

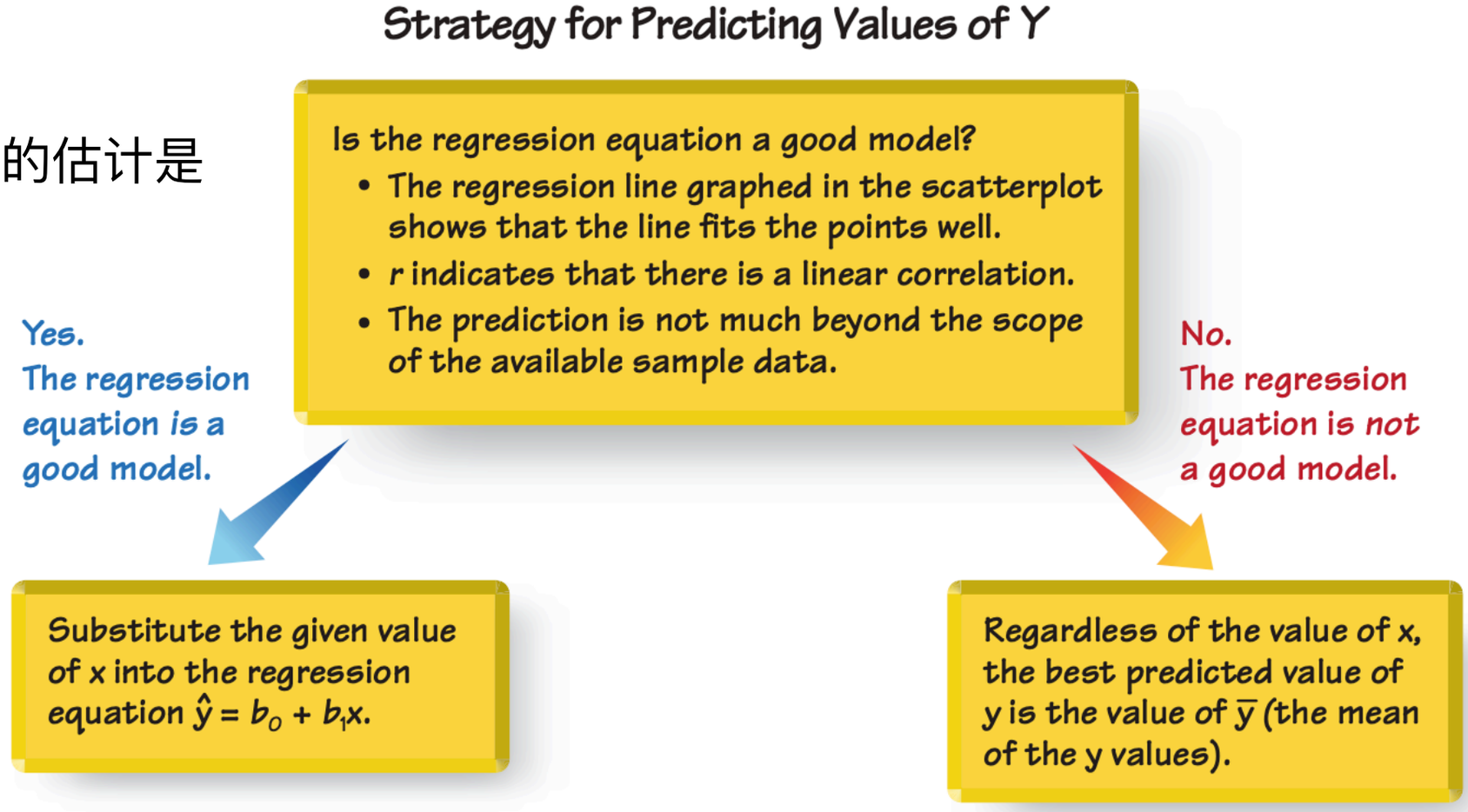
5. 其他：存在时间效应吗?

利用回归的结果进行预测的策略

- 拟合效果好（图、R-square）
- 没有超出样本范围比如便宜x的极值点太远
- 如果拟合效果不好，最好的估计是y-mean

利用回归的结果进行预测的策略

- 拟合效果好（图、R-square）
- 没有超出样本范围比如便宜x的极值点太远
- 如果拟合效果不好，最好的估计是y-mean



多元线性回归

- 回归模型、回归方程
- 最小二乘法
- 解读模型结果
- 综合验证
- 模型比较
- 变量选择

回归模型、回归方程

- 回归模型：$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$
- 回归方程：$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$
- 估计的回归方程：$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$

最小二乘法

$$min\Sigma(y_i - \hat{y}_i)^2$$

解读模型结果

- 判定系数
- 回归系数

线性模型假设的综合验证

- library(gvlma)

多重共线性

- 方差膨胀因子：VIF

模型比较

- ANOVA
- AIC （）

变量选择

- stepAIC()

小结

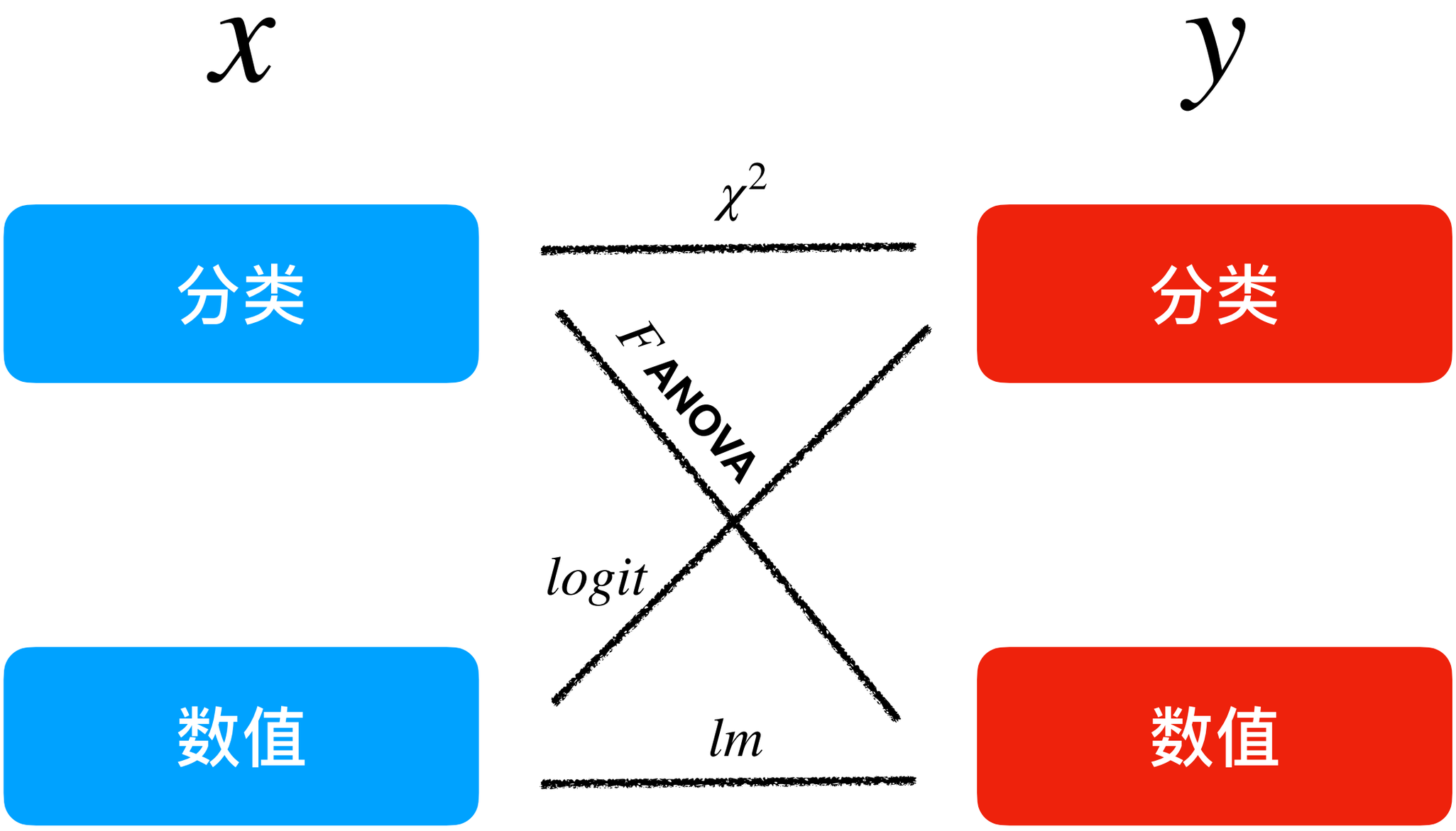
- 先画散点图
- 拟合线性模型
- 模型解读
- 模型评价
- 变量选择、三类点

练习

1. 检验各条线路的均价是否存在显著差异

2.选择除总价（price_ttl)以外的其他其他变量，拟合回归方程并进行解释。

Logistic回归



Logistic回归方程

LOGISTIC REGRESSION EQUATION

$$E(y) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}$$

结果解读

练习

- 针对WE.xlsx数据集，采用logit回归，建立回归方程，对客户流失进行预测。

Wrap-up